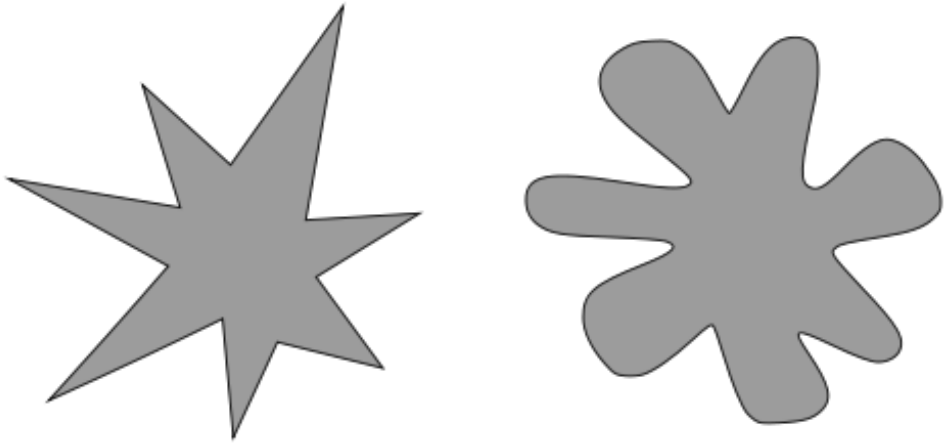
The contribution of sound symbolic evidence to lexically-conditioned phonology

Stephanie S Shih
University of Southern California

NELS 50 | MIT
26 October 2019

Sound symbolism

- Form-meaning correspondences



KIKI **BOUBA**

Sound symbolism

- Often thought to arise from “automatic” factors such as perceptuomotor analogy between phonology and real-world attributes/entities. (e.g., Dingemanse et al. 2015 and references therein)
  - e.g., different frequency resonances in high/low vowels for small/big objects (Ohala 1994)
  - e.g., lengthening of phonological material to express real-world length

Sound symbolism in formal phonology

- The abstraction of form as separate from meaning (or concept) is foundational in modern linguistic study.
- Result: separation of ‘non-arbitrary’ data from the purview of formal theory.

Sound symbolism in formal phonology

- Some recent work has called for bringing sound symbolic patterns back under the umbrella of formal phonological models:
  - On modeling palatalization using Express constraints in OT (Alderete & Kochetov 2016; before OT: Mester & Ito 1989)
  - On sound symbolism in MaxEnt HG (Kawahara et al. 2019)
  - On ideophones in the grammar (e.g., Newman 2001; Dingemanse 2012; and many others)

Sound symbolism in formal phonology

- Some recent work has called for bringing sound symbolic patterns back under the umbrella of formal phonological models:
  - On modeling palatalization using Express constraints in OT (Alderete & Kochetov 2016; before OT: Mester & Ito 1989)
  - On sound symbolism in MaxEnt HG (Kawahara et al. 2019)
  - On ideophones in the grammar (e.g., Newman 2001; Dingemanse 2012; and many others)

- Sound symbolism, in spite of its non-arbitrary roots, is not as extraordinary from "core" phonological patterns as we traditionally believed.

This talk

- Sound symbolic patterns parallel phonological patterns that our “core” phonological models already capture.
- Sound symbolism is constrained by and interacts with linguistic structure.
- Real-world attributes shape the categories that are relevant to phonological patterns.
- Capturing sound symbolic patterns requires some natural extensions to our existing formal models: namely, in how we deal with the categories that are relevant to phonological patterns.

This talk: a roadmap

1. **Setting up the theoretical environment**
   - Lexically-conditioned phonology in MaxEnt grammar
   - A toy illustration

2. **Two sound symbolism case studies, demonstrating "naturalistic" sound symbolic patterns**
   - Cross-linguistic dataset: Pokémon
   - "Real-world" dataset

3. **Theoretical ramifications**
   - Return to the toy illustration

Setting up the theoretical environment.

Lexically-conditioned phonology and a toy illustration

Sound symbolism in formal phonology

- Sound symbolism highlights the interaction between categories and the different phonological patterns that arise between categories.

Sound symbolism in formal phonology

- Sound symbolism highlights the interaction between categories and the different phonological patterns that arise between categories.

- Familiar topic in “core” morphophonology, where different categories often exhibit different phonotactic patterns and phonological behaviours.
  - namely, in **lexically-conditioned phonology**
  - e.g., content words, nouns exhibit greater phonotactic contrasts
    vs. function words, which exhibit greater phonological reductions

(e.g., Selkirk 1984, 1996; Kaisse 1985; Inkelas & Zec 1993; Kelly & Bock 1988; Kelly 1992; Segalowitz & Lane 2000; Bell et al. 2009; Smith 2011, 2016; Shih 2014, 2018)

Lexically-conditioned phonology

Categories relevant for lexically-conditioned phonology:

- morphosyntactic categories
  - content versus function classes (e.g., Kelly & Bock 1988)
  - parts of speech (e.g., Smith 2011, 2016)
- etymological strata (e.g., Itô & Mester 1999, 2003, 2009)
- "semantic" categories
  - ideophones (e.g., Newman 2001; Dingemanse 2012; Rose 2015; Shih & Inkelas 2016)
- extra-linguistically-defined categories
  - sex (e.g., Slater & Feinman 1985; Cutler et al. 1990)
- potentially arbitrary categories
  - arbitrary gender classes
  - other arbitrary categories (e.g., DuBois 1985)

*** This is not an exhaustive list.**

Toy Illustration

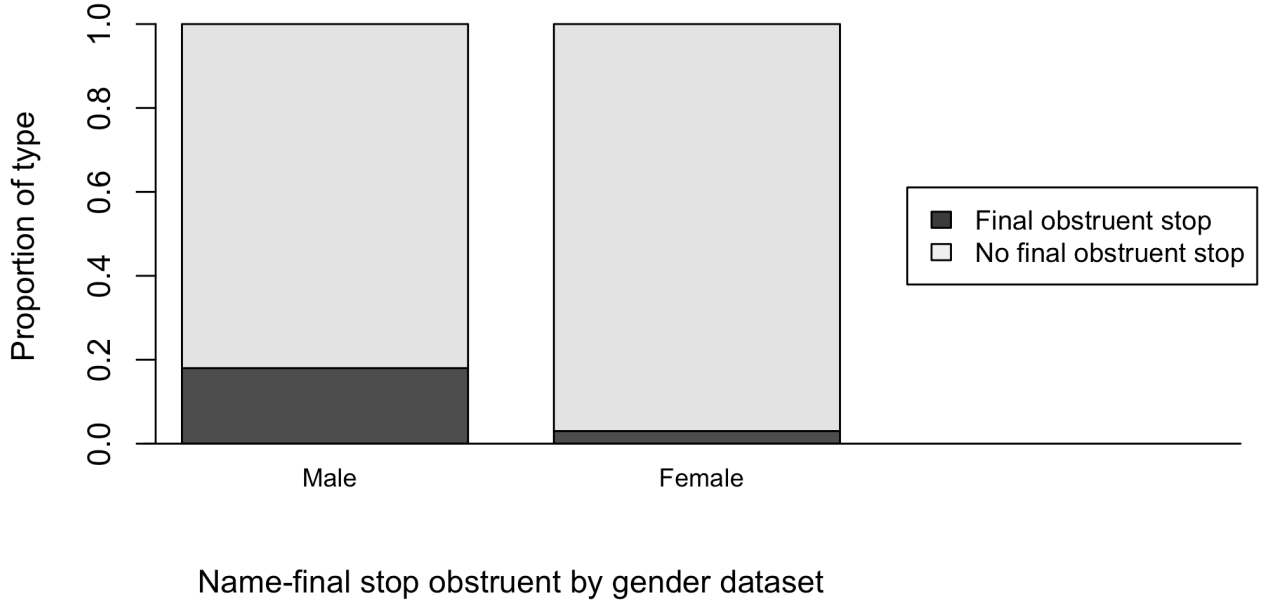
Male vs. Female names

- Data:
  - American English names between 1990–1999 (Social Security Administration)
  - Most frequent 200 male and 200 female names

- Phonotactic differences between name genders were taken from the previous literature. (e.g., Cassidy et al. 1999; Wright et al. 2005)
  - Here: focus on the two strongest predictors of male/female names in the current dataset (as tested with the MaxEntGrammarTool (Hayes et al. 2009)).

Toy Illustration

Male vs. Female names



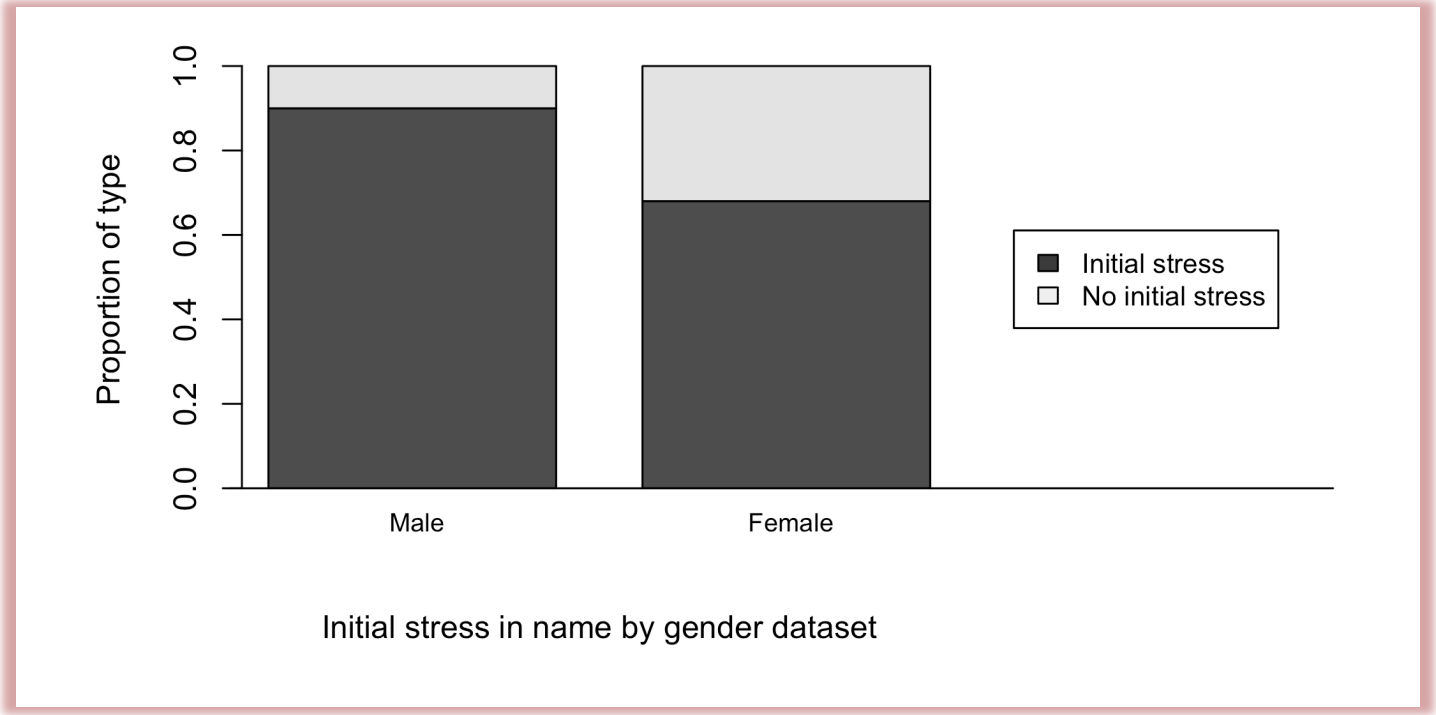
- ***T#**

Female names avoid final stop obstruents more than male names.

♀	♂
Elaine	Albert
…[n]#	…[t]#

Toy Illustration

Male vs. Female names



- **ϕ=TROCH**

Male names are more likely to begin with initial stress (roughly, trochaic).

♀	♂
Elaine	Albert
σ.ˈσ	ˈσ.σ

Lexically-conditioned phonology

Existing approaches:

- lexically-indexed constraints (e.g., Ito & Mester 1995; Pater 2000. 2009; Smith 2001; Alderete 2009)
- strata (e.g., Kiparsky 1982, et seq.)
- cophonologies (e.g., Anttila 2002; Inkelas & Zoll 2005)
- sublexical grammars (e.g., Becker & Gouskova 2016; Allen & Becker 2015)

*** This is not an exhaustive list.**

Lexically-conditioned phonology

Existing approaches:

- lexically-indexed constraints (e.g., Ito & Mester 1995; Pater 2000. 2009; Smith 2001; Alderete 2009)
- strata (e.g., Kiparsky 1982, et seq.)
- cophonologies (e.g., Anttila 2002; Inkelas & Zoll 2005)
- sublexical grammars (e.g., Becker & Gouskova 2016; Allen & Becker 2015)

- **for phonotactic distributions over a lexicon, we'll use an implementation in MaxEnt HG.**

Lexical conditioning in MaxEnt HG

- **Lexically indexed constraints/cophonologies in MaxEnt HG**
  (Shih & Inkelas 2015; see also Albright 2008; Coetzee & Pater 2011)

$$w_1(\mathbb{C}) + w_2(\mathbb{C} \times \mathbb{k}_i) + w_3(\mathbb{C} \times \mathbb{k}_j) + \ldots w_N(\mathbb{C} \times \mathbb{k}_x)$$

Each constraint $\mathbb{C}$ has a weight w_N for every category $\mathbb{k}$.

Lexical conditioning in MaxEnt HG

- **Lexically indexed constraints/cophonologies in MaxEnt HG**
  (Shih & Inkelas 2015; see also Albright 2008; Coetzee & Pater 2011)

$$w_1(\mathbb{C}) + w_2(\mathbb{C} \times \mathbb{k}_i) + w_3(\mathbb{C} \times \mathbb{k}_j) + \ldots w_N(\mathbb{C} \times \mathbb{k}_x)$$

Each constraint $\mathbb{C}$ has a weight w_N for every category $\mathbb{k}$.

- $\sim$ varying slopes in multilevel statistical models
  (here treated as an interaction term; see e.g., Gelman & Hill 2007 for discussion of the equivalency of interaction terms and random slopes in this situation.)

Lexical conditioning in MaxEnt HG

$$w_1(\mathbb{C}) + w_2(\mathbb{C} \times \text{Male}) + w_3(\mathbb{C} \times \text{Female}) + \ldots w_N(\mathbb{C} \times \mathbb{k}_x)$$

‘Base’ grammar = $w_1(\mathbb{C})$

Male ‘cophonology’ = $w_1(\mathbb{C}) + w_2(\mathbb{C} \times \text{Male})$

Female cophonology = $w_1(\mathbb{C}) + w_3(\mathbb{C} \times \text{Female})$

Toy Illustration

Male vs. Female names

- a pseudogrammar with lexically-indexed constraints, for ♂ and ♀

Male Female "Base" grammar

			TROCH♂	*T]#♀	WSP♀	TROCH	*T]#	WSP	FAITH-C	
			3	3	2	1	1	1	2	$\mathcal{H}$
(1) /CV.CVT/ ♂	☞	a. ˈCV.CVT					–1	–1		–2
		b. CV.ˈCVN	–1			–1			–1	–6
		c. ˈCV.CVN						–1	–1	–3

Toy Illustration

Male vs. Female names

- a pseudogrammar with lexically-indexed constraints, for ♂ and ♀

Male Female "Base" grammar

			TROCH♂	*T]#♀	WSP♀	TROCH	*T]#	WSP	FAITH-C	
			3	3	2	1	1	1	2	$\mathcal{H}$
(1) /CV.CVT/ ♂	☞	a. ˈCV.CVT					–1	–1		–2
		b. CV.ˈCVN	–1			–1			–1	–6
		c. ˈCV.CVN						–1	–1	–3
(2) /CV.CVT/ ♀		a. ˈCV.CVT		–1	–1		–1	–1		–7
	☞	b. CV.ˈCVN				–1			–1	–3
		c. ˈCV.CVN			–1			–1	–1	–5

Modeling lexically-conditioned phonology

- So far, the current available approaches to lexically-conditioned phonology appear to work well, even for sound symbolism seen heretofore.
- But, what about naturally-occurring sound symbolic phonological patterns?
  - Turns out, we will need more gradience in our system to capture existing sound symbolic patterns.

- Next two case studies:
  - Demonstrate the range of sound symbolic behaviours in naturalistic datasets.

Case study 1.

Sound symbolism in Pokémon names

Pokémon and the Pokéverse

- Pocket Monsters (aka: Pokémon): 1995-present videogame franchise
  - many video games
  - trading cards
  - TV shows
  - 21 movies
  - a musical
  - viral app

- Goals in game:
  - complete *Pokédex* by collecting (at least one instance) of each Pokémon species
  - train Pokémon to battle. Winning battles increases stats, abilities, Pokémon development so that it is easier to capture new species.

Why Pokémon?

- Challenges to studying cross-linguistic sound symbolism:
  - finding (rather than stumbling upon) sound symbolic correspondences
  - insufficiently large and rich datasets
  - controlled real-world attributes cross-culturally (e.g., Bremner et al. 2013)

- The Pokéverse provides a self-contained universe that allows for many potential sound symbolic correspondences.
- Dataset extends cross-linguistically, while holding non-linguistic factors constant across the universe.

Pokémon dataset

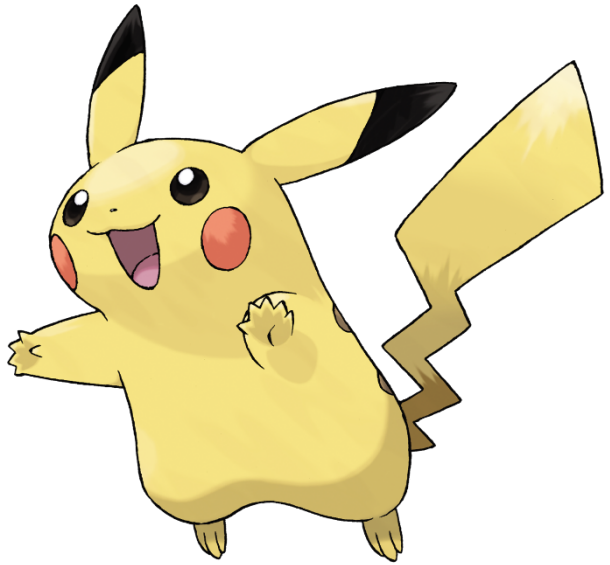
- *N* = 802 Pokémon
- Pokémon attributes and statistics
  - taken from *Sun* and *Moon* versions of the game (2016–2017)
  - source: Bulbapedia, a fan-based encyclopedia of Pokémon data

- Pokémonikers in current case study
  - Japanese
  - English
  - Mandarin/Cantonese Chinese

Pokémonikers

- **Japanese**
  Hitokage — *hi* ‘fire’ + *tokage* ‘lizard’
- **English**
  Charmander — *char* + *salamander*
- **Chinese**
  小火龙 — *小* ‘small’ + *火* ‘fire’ + *龙* ‘dragon’
- **Korean**
  파이리 *fairit* — *파이/어* ‘fire’ (translit) + *꼬리* ‘tail’
- **Russian**
  Чармандер — [ʧer'mander]

Pokémonikers

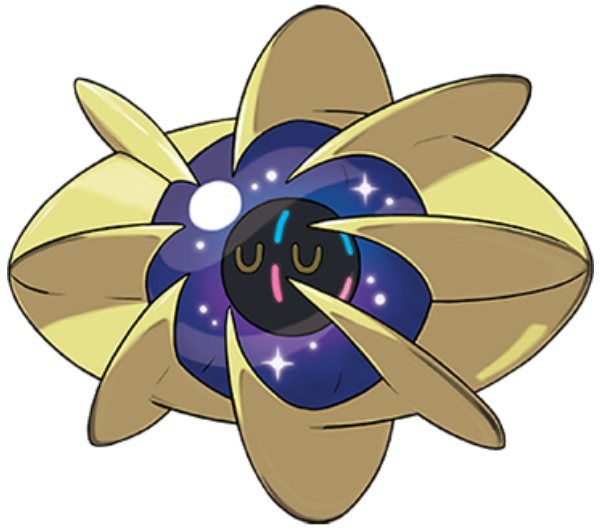


- **Japanese**
  Pikachu — *pikapika* ‘sparkle’ + *chuuchuu* ‘squeaking’
- **English**
  Pikachu — [pi.ka.ʧu]
- **Chinese**
  皮卡丘 — [pʰi^{35} kʰa^{21} tɕʰiou^{55}]
- **Korean**
  피카츄 — [pʰi.kʰa.cʰju]
- **Russian**
  Пикачу — [pʲi.ka.ʧu]

Pokémon attributes

- Appearance
  - Weight
  - Height

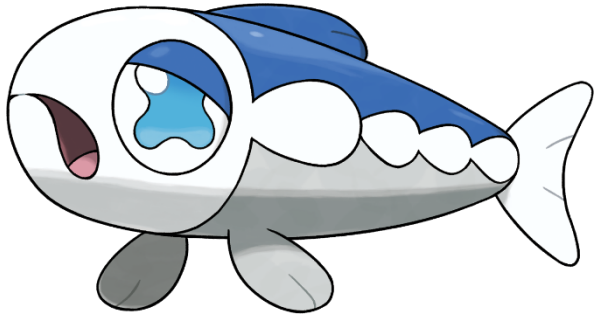
Cosmog
0.1 kg



Cosmoem
999.9 kg

Pokémon attributes

- Appearance
  - Weight
  - Height
- Power (Total power)
  - Hit Points (HP)
  - Attack
  - Defense
  - Special Attack
  - Special Defense
  - Speed



Wishiwashi
Total Power = 175

Arceus
Total Power = 720

Pokémon attributes

- Appearance
  - Weight
  - Height
- Performance (Stats)
  - Health Points (HP)
  - Attack
  - Defense
  - Special Attack
  - Special Defense
  - Speed
- Evolutionary stage
  - Baby
  - Stages 1,2,3
  - Legendary

***Abra* (Stage 1)**

***Kadabra* (Stage 2)**

***Alakazam* (Stage 3)**

Pokémon attributes

- Appearance
  - Weight
  - Height
- Performance (Stats)
  - Health Points (HP)
  - Attack
  - Defense
  - Special Attack
  - Special Defense
  - Speed
- Evolutionary stage
  - Baby
  - Stages 1,2,3
  - Legendary
- Gender
  - % Male/Female

Braviary
100% M, 0% F

Bunnelby
50% M, 50% F

Blissey
0% M, 100% F

Phonological features investigated

	Japanese	English	Mandarin	Cantonese
Prosodic factors	moras syllables	segments syllables	segments syllables low/high tone contour/level tone tones 0–4	segments syllables low tone rising tone tones 1–6

Phonological features investigated

	Japanese	English	Mandarin	Cantonese
Prosodic factors	moras syllables	segments syllables	segments syllables low/high tone contour/level tone tones 0–4	segments syllables low tone rising tone tones 1–6
Vowel quality	height backness	height backness	height backness diphthong	height backness

Phonological features investigated

	Japanese	English	Mandarin	Cantonese
Prosodic factors	moras syllables	segments syllables	segments syllables low/high tone contour/level tone tones 0–4	segments syllables low tone rising tone tones 1–6
Vowel quality	height backness	height backness	height backness diphthong	height backness
Consonant quality	voicing [+v/-v] sonorant labial palatal velar	voicing [+v/-v] sonorant labial alveolar velar	voicing [+v/-v] sonorant labial retroflex velar nasal palatal	voicing [+v/-v] final p/t/k labial velar nasal

Phonological features investigated

	Japanese	English	Mandarin	Cantonese
Prosodic factors	moras syllables	segments syllables	segments syllables low/high tone contour/level tone tones 0–4	segments syllables low tone rising tone tones 1–6
Vowel quality	height backness	height backness	height backness diphthong	height backness
Consonant quality	voicing [+v/-v] sonorant labial palatal velar	voicing [+v/-v] sonorant labial alveolar velar	voicing [+v/-v] sonorant labial retroflex velar nasal palatal	voicing [+v/-v] final p/t/k labial velar nasal
Morpho			reduplication	reduplication

Results at a glance

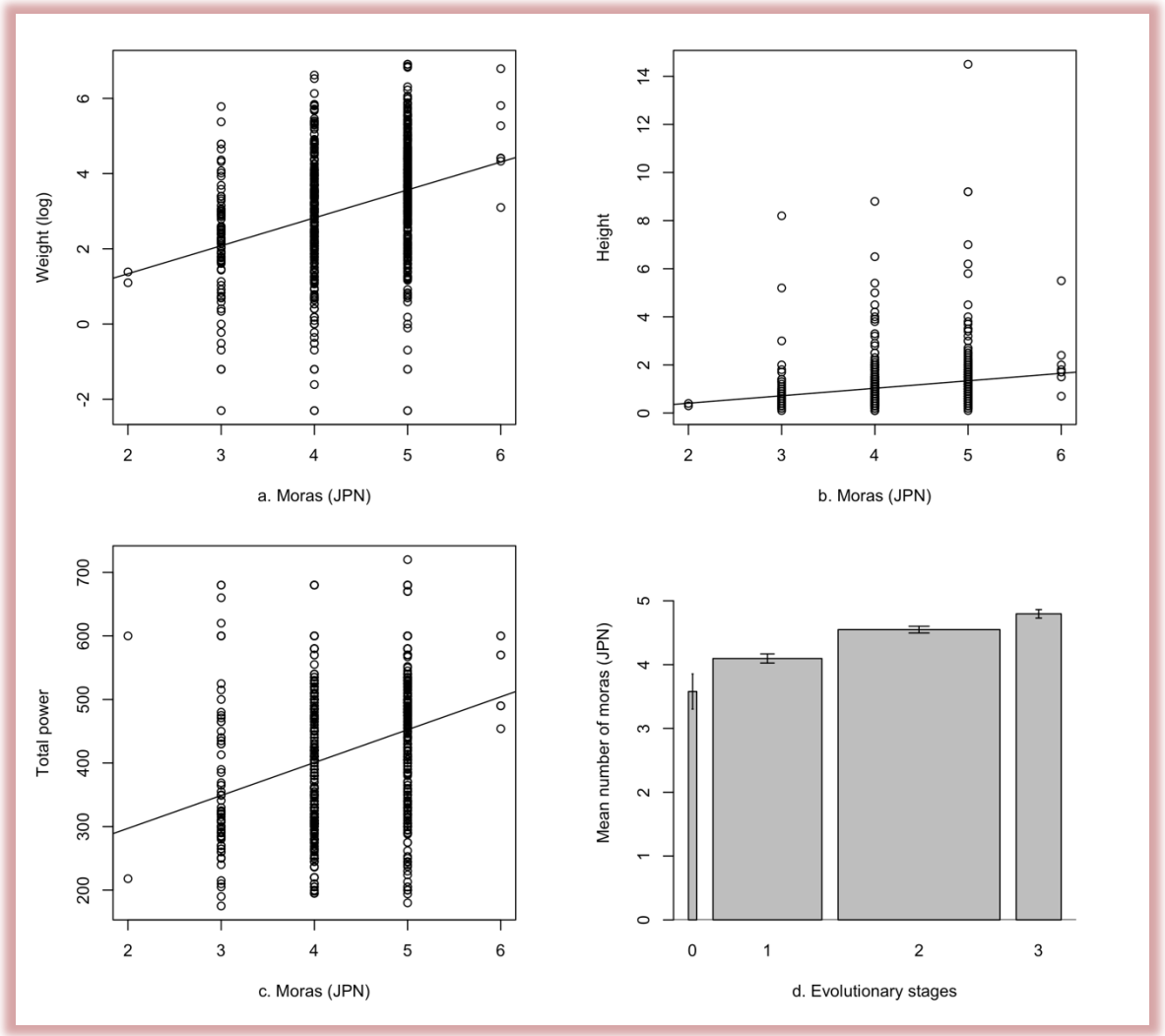
	Japanese	English	Mandarin	Cantonese (HK)
Appearance Weight				
Height				
Power				
Stage				
Gender (% Male)				

How to read the results

+	positive correlation
–	negative correlation
bold	significant across the entire ensemble set of models
regular	significant or trending across ensemble set of models

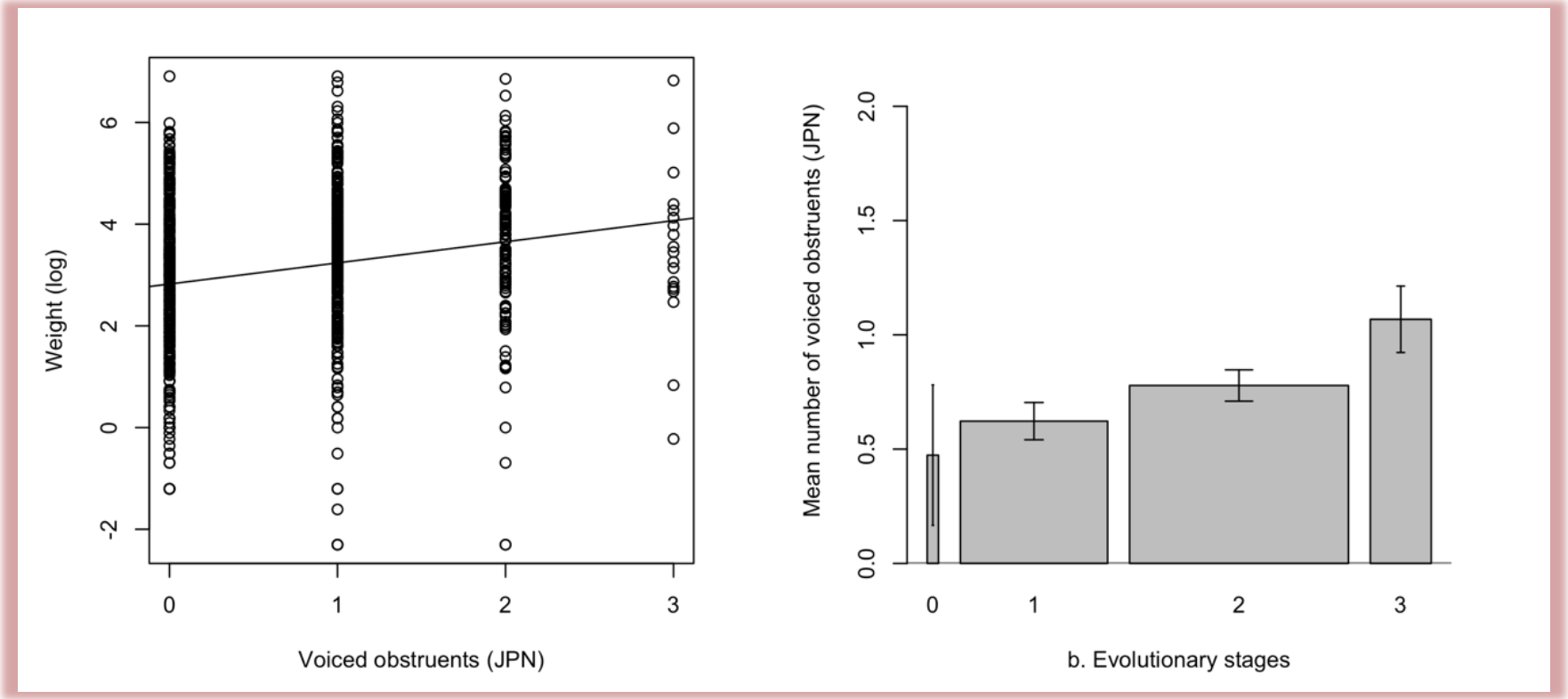
Results: Japanese

	Japanese
Appearance Weight	**+ moras** **+ voiced obstr** – labial C – palatal C
Height	**+ mora** **– labial C**
Power	**+ mora** **– labial C**
Stage	**+ mora** + voiced obstr – labial C
Gender (% Male)	**– sonorant C**



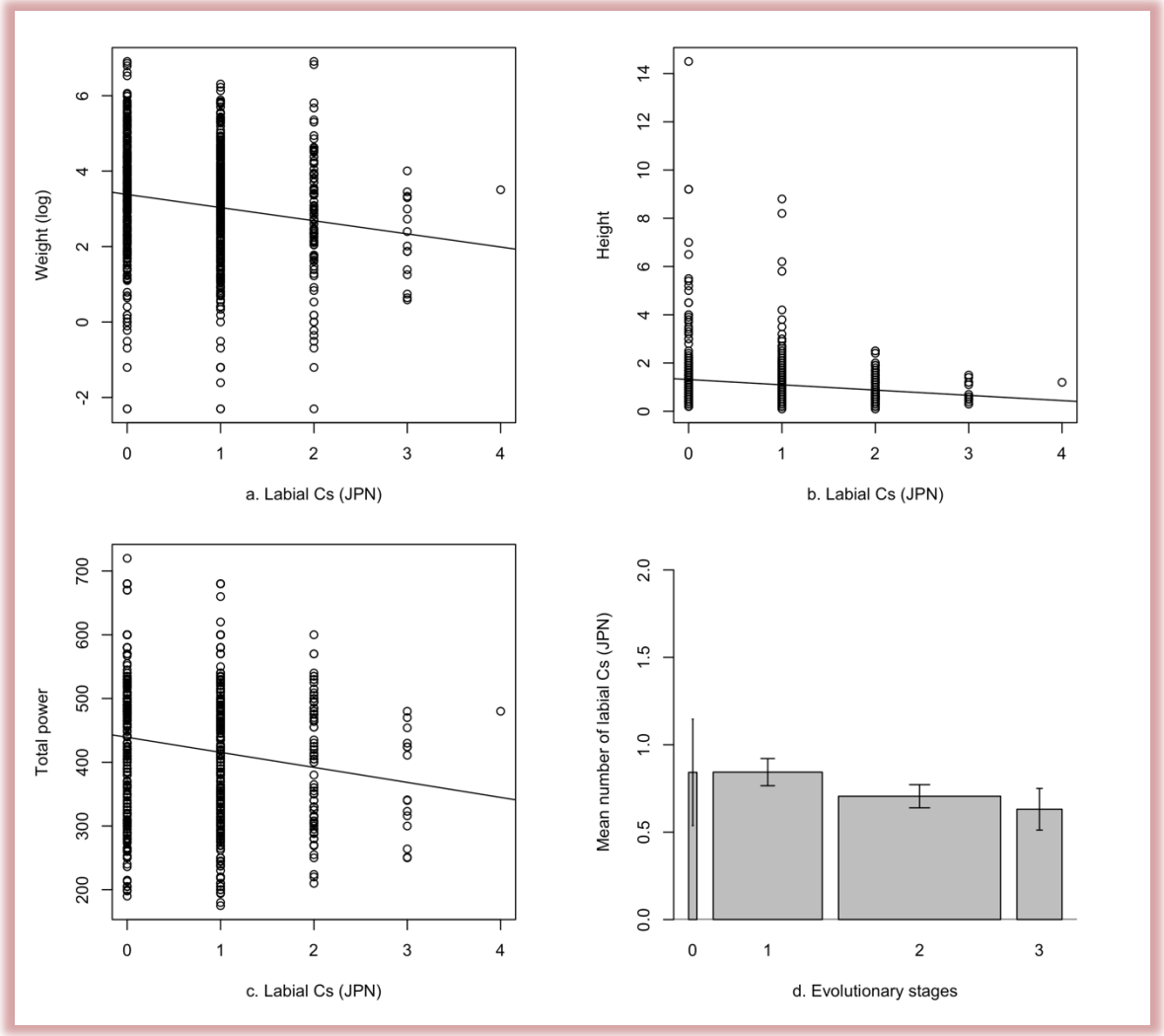
Results: Japanese

	Japanese
Appearance Weight	**+ moras** **+ voiced obstr** – labial C – palatal C
Height	**+ mora** **– labial C**
Power	**+ mora** **– labial C**
Stage	**+ mora** + voiced obstr – labial C
Gender (% Male)	**– sonorant C**



Results: Japanese

	Japanese
Appearance Weight	**+ moras** **+ voiced obstr** – labial C – palatal C
Height	**+ mora** **– labial C**
Power	**+ mora** **– labial C**
Stage	**+ mora** + voiced obstr – labial C
Gender (% Male)	**– sonorant C**

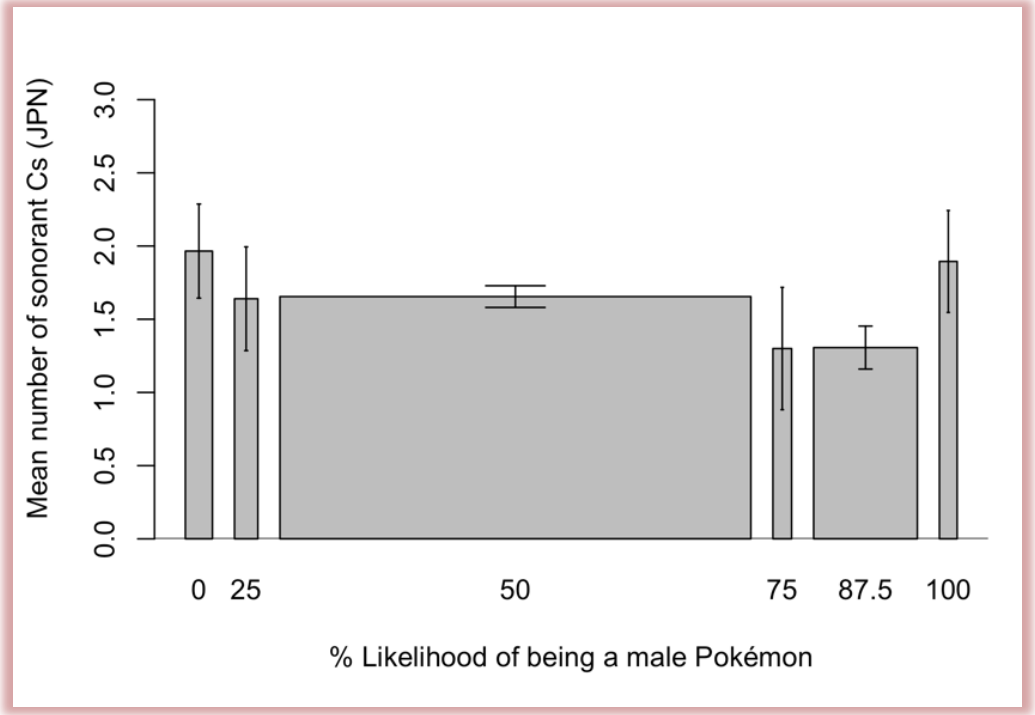


Labial sound symbolism in Japanese

- Japanese has a propensity for labial sounds to associate with baby items (e.g., diaper brands).
  (Kumagai & Kawahara 2017)

Results: Japanese

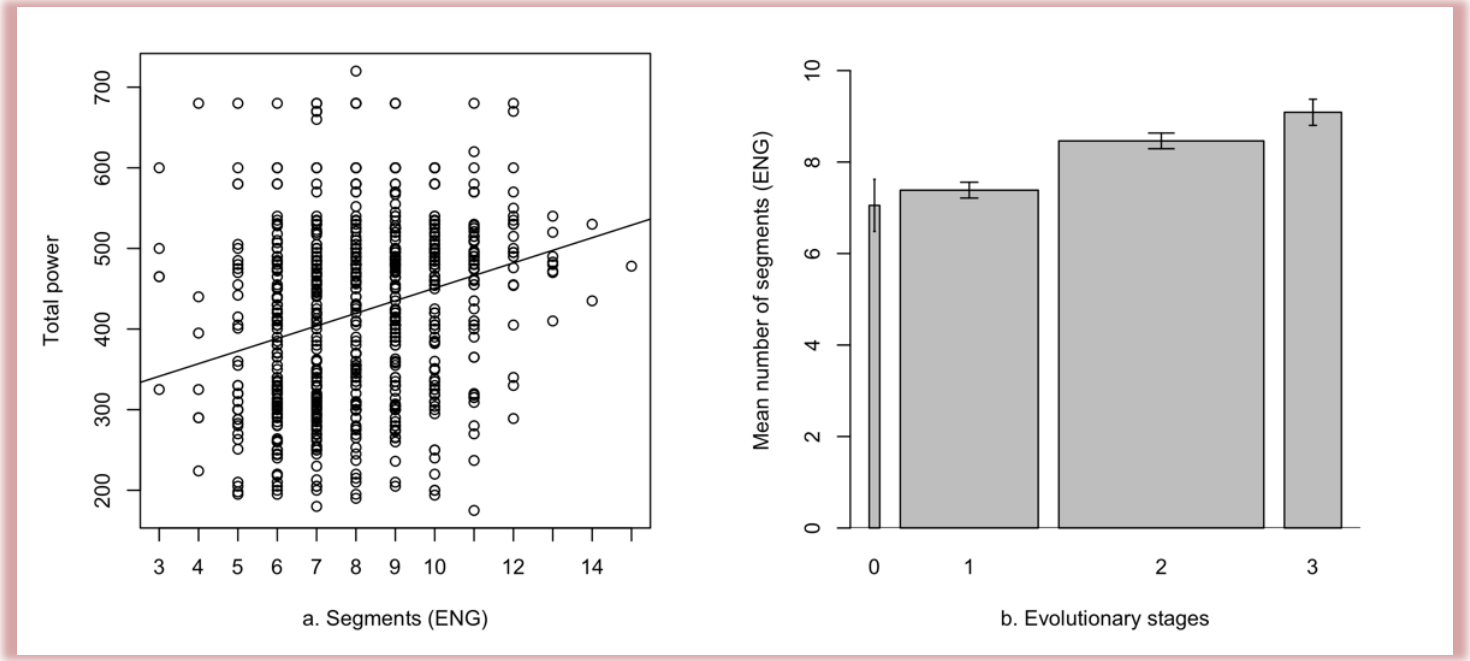
	Japanese
Appearance	**+ moras**
Weight	**+ voiced obstr**
	– labial C
	– palatal C
Height	**+ mora**
	– labial C
Power	**+ mora**
	– labial C
Stage	**+ mora**
	+ voiced obstr
	– labial C
Gender (% Male)	**– sonorant C**



- Cf: of the top 50 male/50 female names in Japanese, female names are significantly more likely to include sonorant consonants than male names. (Shinohara & Kawahara 2013)

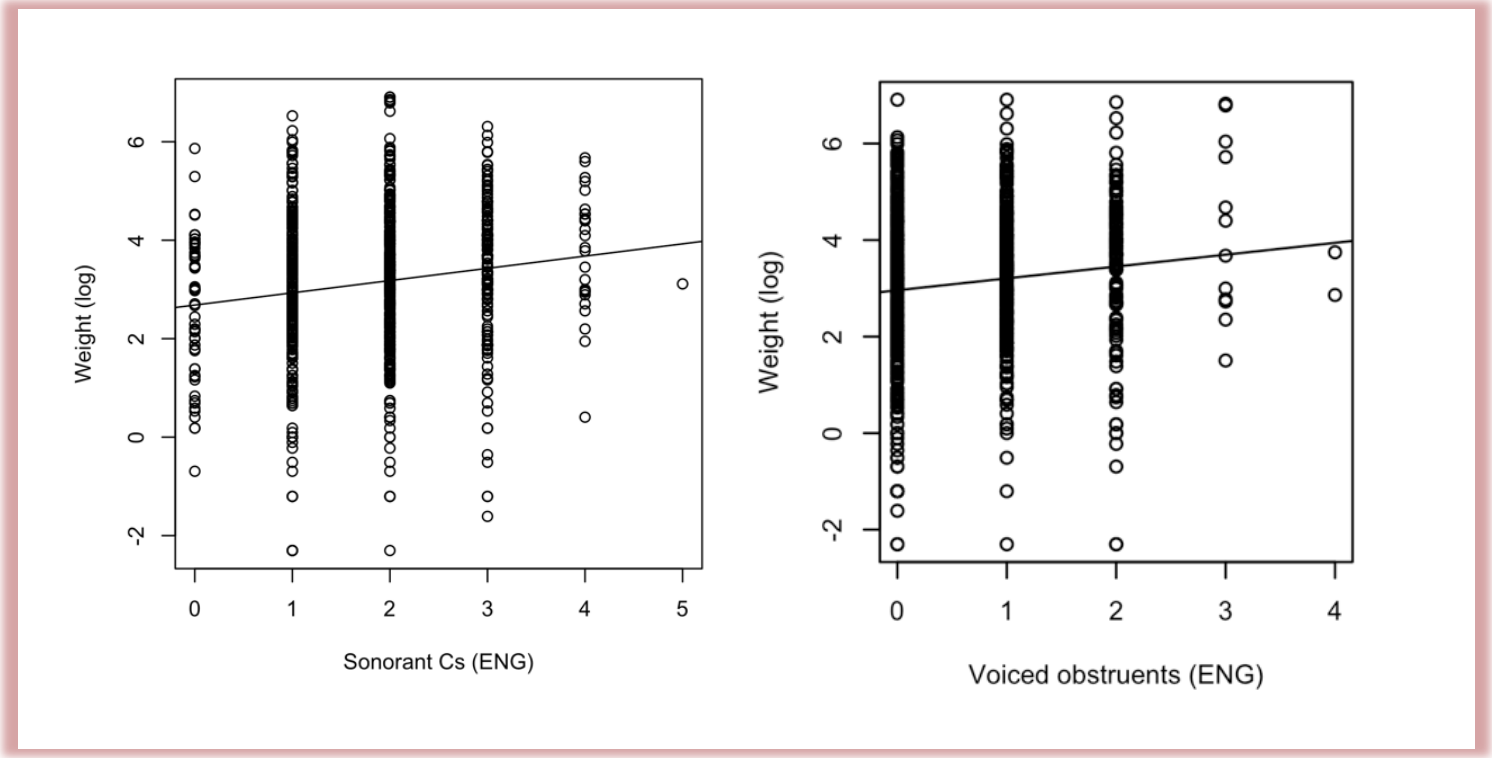
Results: English

	Japanese	English
Appearance Weight	+ moras	**+ sonorant C**
	+ voiced obstr	+ voiced obstr
	– labial C	– labial C
	– palatal C	+ low V
Height	+ mora	**– labial C**
	– labial C	+ low V
Power	+ mora	**+ segments**
	– labial C	**– labial C**
		+ low V
Stage	+ mora	**+ segments**
	+ voiced obstr	
	– labial C	
Gender (% Male)	– sonorant C	**+ back V**



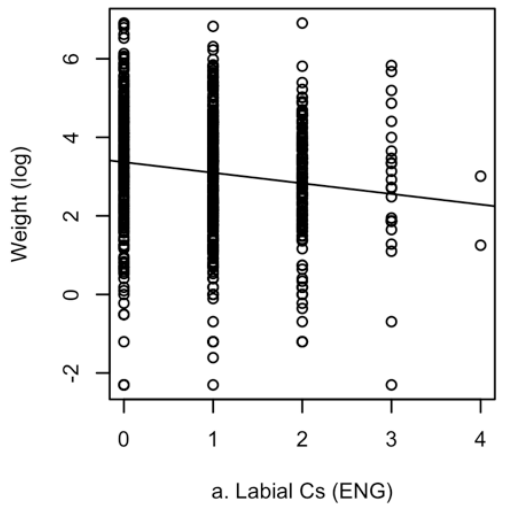
Results: English

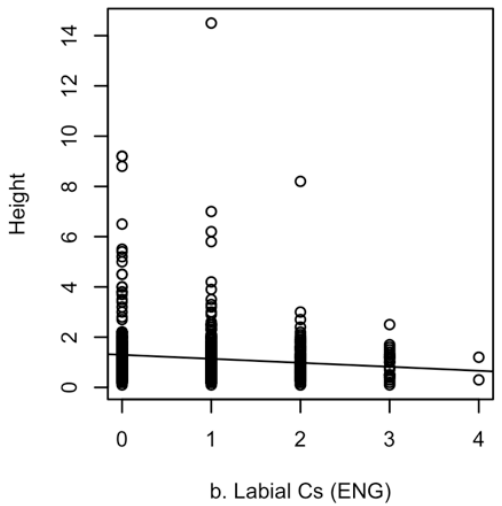
	Japanese	English
Appearance Weight	+ moras	**+ sonorant C**
	+ voiced obstr	+ voiced obstr
	– labial C	– labial C
	– palatal C	+ low V
Height	+ mora	**– labial C**
	– labial C	+ low V
Power	+ mora	**+ segments**
	– labial C	**– labial C**
		+ low V
Stage	+ mora	**+ segments**
	+ voiced obstr	
	– labial C	
Gender (% Male)	– sonorant C	**+ back V**

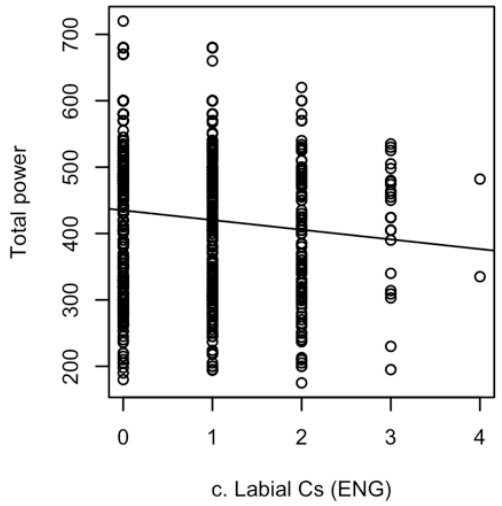


Results: English

	Japanese	English
Appearance Weight	+ moras	**+ sonorant C**
	+ voiced obstr	+ voiced obstr
	– labial C	– labial C
	– palatal C	+ low V
Height	+ mora	**– labial C**
	– labial C	+ low V
Power	+ mora	**+ segments**
	– labial C	**– labial C**
		+ low V
Stage	+ mora	**+ segments**
	+ voiced obstr	
	– labial C	
Gender (% Male)	– sonorant C	**+ back V**

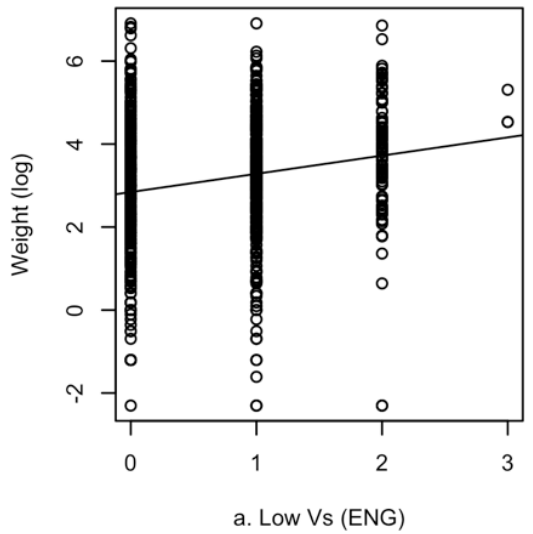


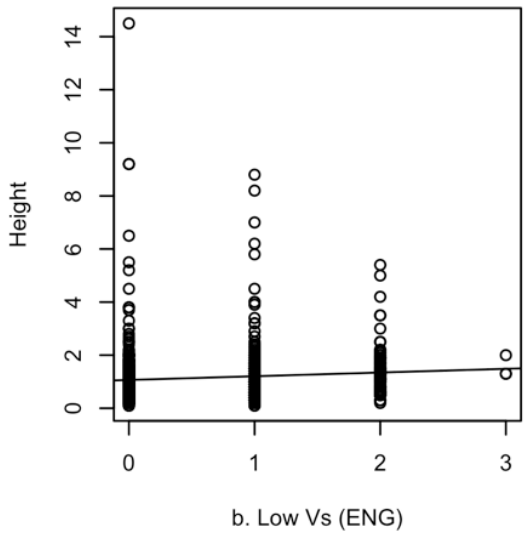


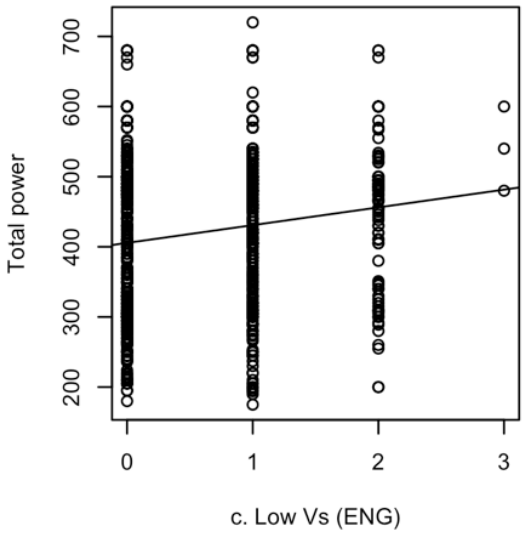


Results: English

	Japanese	English
Appearance Weight	+ moras	**+ sonorant C**
	+ voiced obstr	+ voiced obstr
	– labial C	– labial C
	– palatal C	+ low V
Height	+ mora	**– labial C**
	– labial C	+ low V
Power	+ mora	**+ segments**
	– labial C	**– labial C**
		+ low V
Stage	+ mora	**+ segments**
	+ voiced obstr	
	– labial C	
Gender (% Male)	– sonorant C	**+ back V**

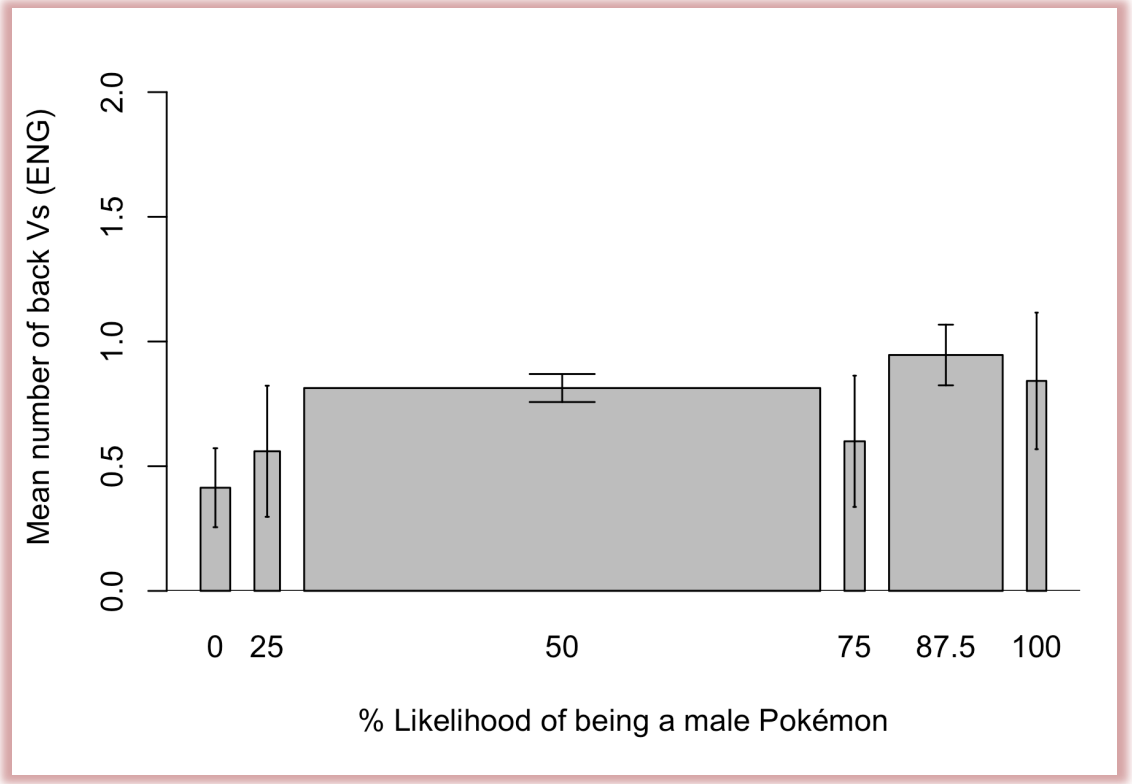






Results: English

	Japanese	English
Appearance Weight	+ moras	**+ sonorant C**
	+ voiced obstr	+ voiced obstr
	– labial C	– labial C
	– palatal C	+ low V
Height	+ mora	**– labial C**
	– labial C	+ low V
Power	+ mora	**+ segments**
	– labial C	**– labial C**
		+ low V
Stage	+ mora	**+ segments**
	+ voiced obstr	
	– labial C	
Gender (% Male)	– sonorant C	**+ back V**



- Cf: of the top 100 male/100 female names in English (1918–2017), there are significantly more back vowels in male names than in female names (χ^2=10.279, p=0.001).

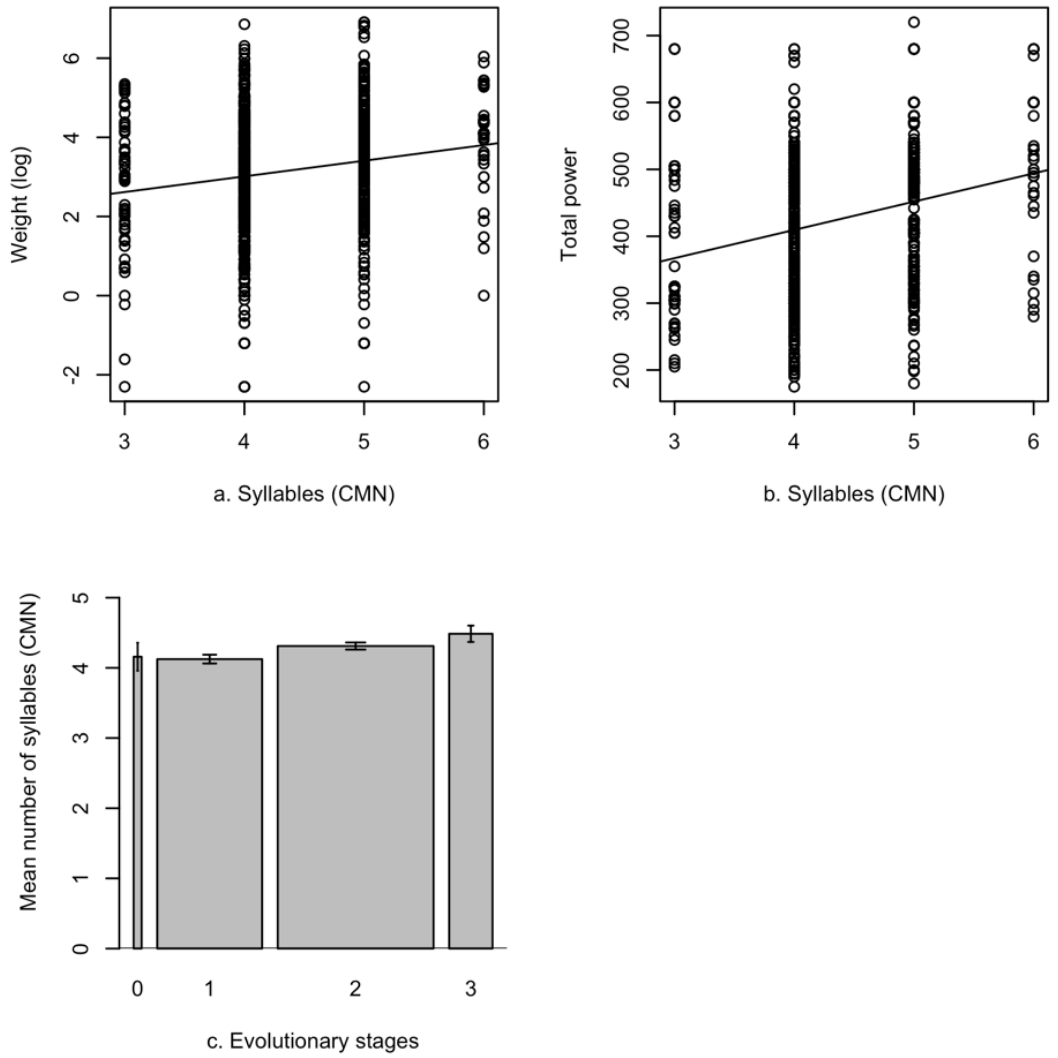
Results: Chinese

	Mandarin	Cantonese (HK)
Appearance	**+ syllables**	**+ syllables**
Weight	**– reduplication**	**– reduplication**
		– non-high V
Height	**– reduplication**	**– reduplication**
		– 2nd tone
Power	**+ syllables**	**– reduplication**
	– reduplication	**– 2nd tone**
	+ front V	
	+ 4th tone	
Stage	**+ syllables**	**+ syllables**
	– reduplication	**– reduplication**
		– 2nd tone
Gender	– reduplication	– reduplication
(% Male)	+ low tones (3rd & 4th)	+ 6th tone

- Unlike English and Japanese, sound symbolic correspondences are mostly prosodic (i.e., non-segmental) or morphophonological in Mandarin and Cantonese.

Results: Chinese

	Mandarin	Cantonese (HK)
Appearance	**+ syllables**	**+ syllables**
Weight	**– reduplication**	**– reduplication**
		– non-high V
Height	**– reduplication**	**– reduplication**
		– 2nd tone
Power	**+ syllables**	**– reduplication**
	– reduplication	**– 2nd tone**
	+ front V	
	+ 4th tone	
Stage	**+ syllables**	**+ syllables**
	– reduplication	**– reduplication**
		– 2nd tone
Gender (% Male)	– reduplication	– reduplication
	+ low tones (3rd & 4th)	+ 6th tone



Results: Chinese

	Mandarin	Cantonese (HK)
Appearance Weight	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** – non-high V
Height	**– reduplication**	**– reduplication** – 2nd tone
Power	**+ syllables** **– reduplication** **+ front V** + 4th tone	**– reduplication** **– 2nd tone**
Stage	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** **– 2nd tone**
Gender (% Male)	– reduplication + low tones (3rd & 4th)	– reduplication + 6th tone

- Reduplication in Chinese is used as a diminuitive formation.

狗 ’dog’
狗狗 ‘doggy’

Results: Chinese

	Mandarin	Cantonese (HK)
Appearance Weight	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** – non-high V
Height	**– reduplication**	**– reduplication** – 2nd tone
Power	**+ syllables** **– reduplication** **+ front V** + 4th tone	**– reduplication** **– 2nd tone**
Stage	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** **– 2nd tone**
Gender (% Male)	– reduplication + low tones (3rd & 4th)	– reduplication + 6th tone

- Reduplication in Chinese is associated with female names.
  (Starr et al. 2018)

♂: 亮 [liang^{51}]

♀: 珊珊 [ʂan^{55} ʂan^{55}]

Results: Chinese

	Mandarin	Cantonese (HK)
Appearance Weight	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** – non-high V
Height	**– reduplication**	**– reduplication** – 2nd tone
Power	**+ syllables** **– reduplication** **+ front V** + 4th tone	**– reduplication** **– 2nd tone**
Stage	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** **– 2nd tone**
Gender (% Male)	– reduplication + low tones (3rd & 4th)	– reduplication + 6th tone

Baby stage
Cleffa
皮宝宝 *píbǎobǎo*

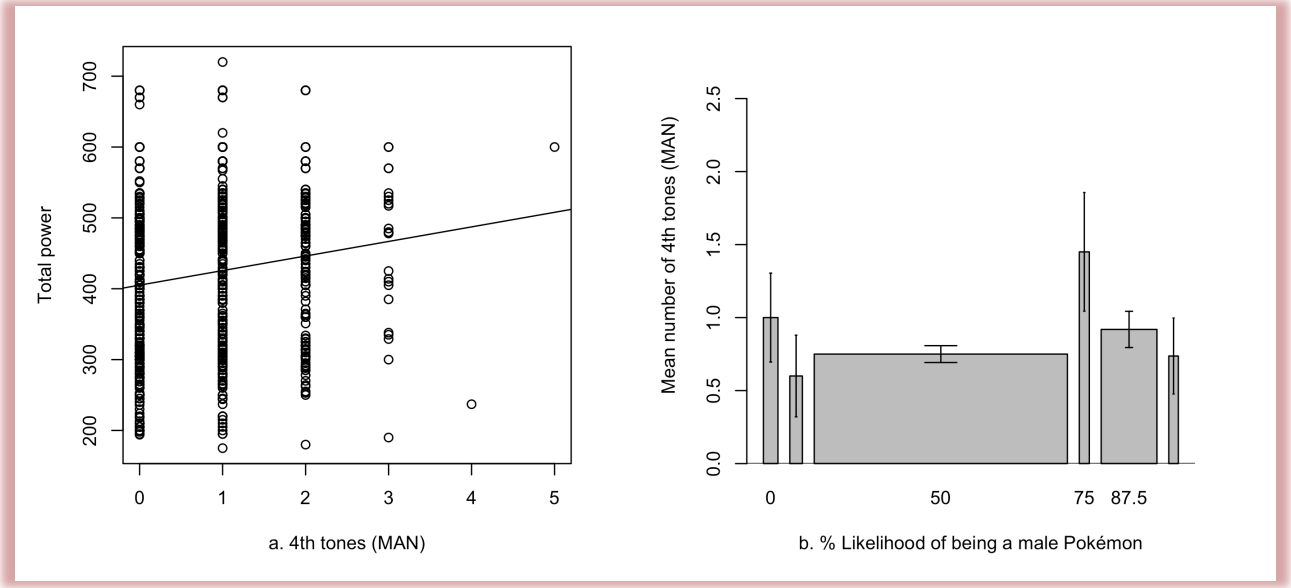
Stage 1
Clefairy
皮皮 *pípí*

Stage 2
Clefable
皮可西 *píkěxī*

Results: Chinese

	Mandarin	Cantonese (HK)
Appearance Weight	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** – non-high V
Height	**– reduplication**	**– reduplication** **– 2nd tone**
Power	**+ syllables** **– reduplication** **+ front V** + 4th tone	**– reduplication** **– 2nd tone**
Stage	**+ syllables** **– reduplication**	**+ syllables** **– reduplication** **– 2nd tone**
Gender (% Male)	– reduplication + low tones (3rd & 4th)	– reduplication + 6th tone

Mandarin 4th tone = high falling

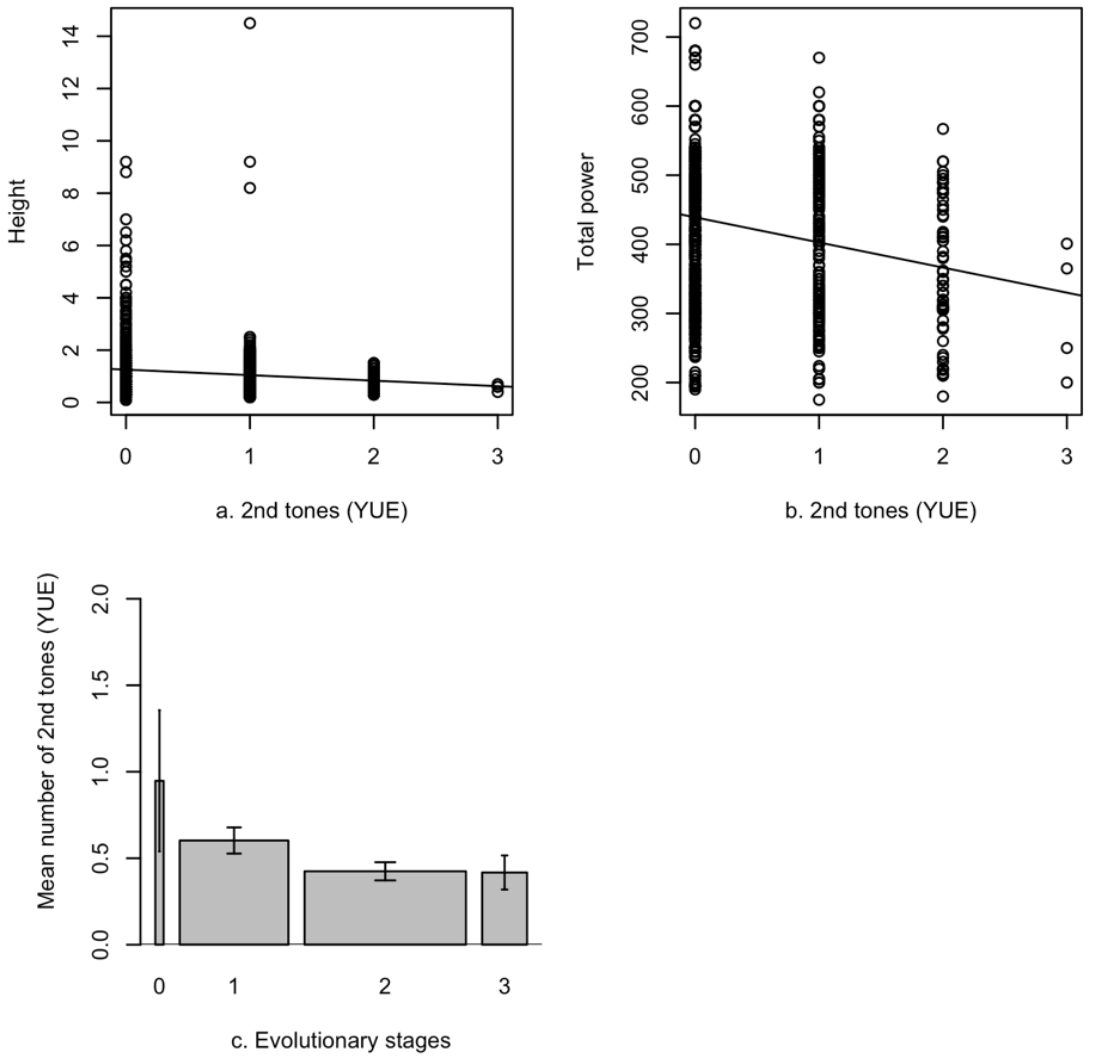


- Cf: Mandarin 4th tone is known to be associated with male names (Starr et al. 2018).

Results: Chinese

	Mandarin	Cantonese (HK)
Appearance	**+ syllables**	**+ syllables**
Weight	**– reduplication**	**– reduplication**
		– non-high V
Height	**– reduplication**	**– reduplication**
		– 2nd tone
Power	**+ syllables**	**– reduplication**
	– reduplication	**– 2nd tone**
	+ front V	
	+ 4th tone	
Stage	**+ syllables**	**+ syllables**
	– reduplication	**– reduplication**
		– 2nd tone
Gender	– reduplication	– reduplication
(% Male)	+ low tones (3rd & 4th)	+ 6th tone

Cantonese 2nd tone = high rising



Cross-linguistic sound symbolic features

- Length of word is possibly the most common phonological correlate to many Pokémon attributes.

- Does this mean that length is a truly cross-linguistic universal sound symbolic effect?
  - e.g., Hinton et al. 1994; Dingemanse et al. 2015 suggest that it's one of the most common cross-linguistic sound symbolic features.

Cross-linguistic sound symbolic features

- Length of word is possibly the most common phonological correlate to many Pokémon attributes.

- Does this mean that length is a truly cross-linguistic universal sound symbolic effect?
  - e.g., Hinton et al. 1994; Dingemanse et al. 2015 suggest that it's one of the most common cross-linguistic sound symbolic features.
- We don't find, in the Pokémon data, any opposite effect—i.e., that in personal names, shorter names are often "stronger."
  - e.g., hypothetical: *Blom* vs. *Teensywinkles*

Cross-linguistic sound symbolic features

- Can length symbolism be explained by morphology?
- I.e., additive morphology is much more common than subtractive, so adding more phonological material is the easiest way to “evolve” a Pokémon.
  - e.g.,

	Subtractive	**Additive**
Stage 1	*Koratta*	*Giaru*
Stage 2	*Ratta*	*Gigiaru*
Stage 3		*Gigigiaru*

Cross-linguistic sound symbolic features

- Controlling for stage:
  - Japanese
    - Name length is still significantly correlated with weight and power.
  - English
    - For weight, name length effect disappears.
    - For power, name length effect still trends, and barely reaches significance.
  - Chinese
    - For weight, name length effect disappears.
    - For power, name length effect trends.

- Word length sound symbolism most active in Japanese.
- In other languages, it's linked primarily to evolution, and likely, the morphological ease with which to evolve characters by addition rather than subtraction.

Cross-linguistic sound symbolic features

- Even length of word is subject to language-specific structural differences:
  - The most effective phonological correlate of word length varies by language.

Psyduck
kodakku

Golduck
gorudakku
*gordakku

Language-specific features

- We also find a number of language-specific sound symbolisms → not all symbolisms are equally cross-linguistic.

Language-specific features

- Both Japanese and English feature two sound-meaning correlations for consonants:
  - More voiced obstruents ⇒ Heavier Pokémon
  - More labial consonants ⇒ Lighter, smaller, less powerful Pokémon

- The Japanese correlations are a bit stronger than the English ones (and they also correlate with evolutionary stage in Japanese).

Language-specific features

- Looking at a subset of data where English and Japanese names are *not* phonologically related (*n*=614):

- The voiced obstruent correlation weakens but remains significant.
  - See Kawahara & Kumagai (to appear): experimental evidence that English speakers have some association of voiced obstruents with evolution, but not as strongly as Japanese speakers. (see also Iwasaki et al. 2007 for a similar, non-Pokémon result)
- The labial consonant correlation weakens and drops out of significance.

- Some effects can be language specific, perhaps learned from experience elsewhere in the language of sound-meaning correspondences.

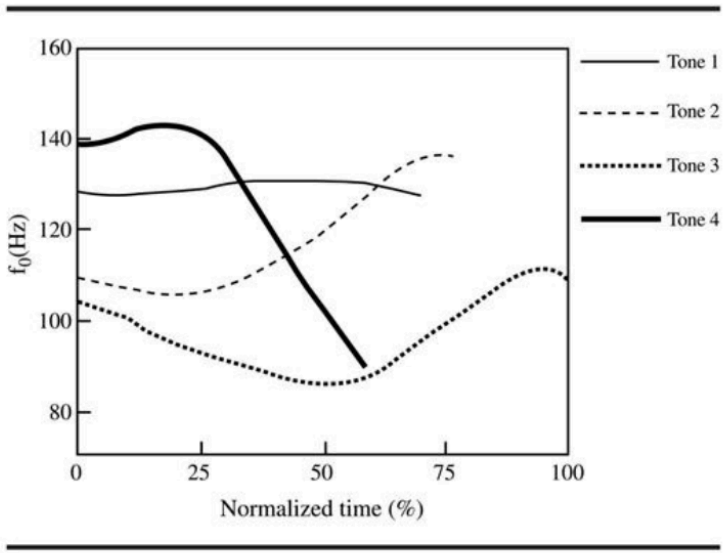
Language-specific features

- For pitch, the usual perceptuomotor analogy expectation is that high pitch should correlate with smaller objects (e.g., Ohala 1984).
- In Pokémon, we find tone-meaning correlations between
  - Mandarin: 4th tone, high-falling
  - Cantonese: 2nd tone, high-rising
  - not 1st tone in both languages, which is H.

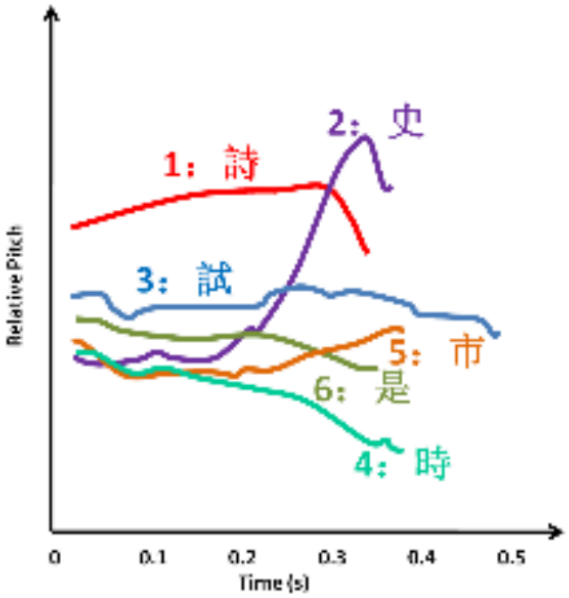
Language-specific features

- Mandarin 4th tone and Cantonese 2nd tone do reach the highest f0 peaks compared to other tones in the respective inventories, but are not level high all the way through.
- They also feature the steepest pitch changes.

Mandarin
(Wong et al. 2005)



Cantonese
(Francis et al. 2008)



Language-specific features

- Mandarin 4th tone and Cantonese 2nd tone have language-specific tone-meaning correspondences beyond Pokémon.
  - Mandarin: words that mean big often have 4th tone.
    - e.g., 大 [ta^{51}]
  - Cantonese: 2nd tone is associated with diminutive reduplication
    - e.g., [je:^{21}] ‘old man, paternal grandfather’ → [je:^{21} je:^{25}] ‘paternal grandfather (with endearment)’

Language-specific features

- Mandarin 4th tone and Cantonese 2nd tone have language-specific tone-meaning correspondences beyond Pokémon.
  - Mandarin: words that mean big often have 4th tone.
    - e.g., 大 [ta^{51}]
  - Cantonese: 2nd tone is associated with diminutive reduplication
    - e.g., [je:^{21}] ‘old man, paternal grandfather’ → [je:^{21} je:^{25}] ‘paternal grandfather (with endearment)’
- Sound symbolic associations in speakers’ lexicons may already predispose them to certain sound/meaning correspondences.

Language-specific features

- Sound symbolism in the Japanese Pokémon names tend to have stronger effect sizes/capture more variance than the sound symbolisms in other languages in this study.
- One possible explanation: Pokémon are ‘native’ to Japan, and there is more sensitivity from the game designers to choosing sound-symbolic names in a familiar language.

Nyoromo (Japanese)
nyoronyoro ‘sound of slithering’ + *kodomo* ‘child’

Poliwag (English)
Possibly from *polliwog* ‘tadpole’

蚊香蝌蚪 ***Wénxiāngkēdǒu*** (Mandarin)
蚊香 ‘mosquito coil’ + 蝌蚪 ‘tadpole’

Language-specific features

- Sound symbolism in the Japanese Pokémon names tend to have stronger effect sizes/capture more variance than the sound symbolisms in other languages in this study.
- Another possibility: stems from many of the Japanese Pokémon names being based on the rich ideophone lexicon.
  - happens more rarely for the English names.
  - Chinese Pokémonikers prioritize semantic content over phonological.

Nyoromo (Japanese)
nyoronyoro 'sound of slithering' + *kodomo* 'child'

Poliwag (English)
Possibly from *polliwog* 'tadpole'

蚊香蝌蚪 ***Wénxiāngkēdǒu*** (Mandarin)
蚊香 'mosquito coil' + 蝌蚪 'tadpole'

Language-specific features

- Sound symbolism in the Japanese Pokémon names tend to have stronger effect sizes/capture more variance than the sound symbolisms in other languages in this study.
- Another possibility: stems from many of the Japanese Pokémon names being based on the rich ideophone lexicon.
  - happens more rarely for the English names.
  - Chinese Pokémonikers prioritize semantic content over phonological.

Muchul (Japanese)
muchuu 'daze' + *chu*
'sound of kiss'

Smoochum (English)
smooch + 'em

迷唇娃 ***Míchúnwá*** (Mandarin)
'bewildering lip doll'

뽀뽀라 [p'op'ora] (Korean)
p'op'o 'kiss'

“Real-world” attributes

- Pokémon attributes that are more closely and robustly associated with sound symbolic patterns are the ones that are more important to achieving game-specific goals:

Power, Stage ≫
Appearance (Weight, Height) ≫
Gender

“Real-world” attributes

- Pokémon attributes that are more closely and robustly associated with sound symbolic patterns are the ones that are more important to achieving game-specific goals:

  Power, Stage ≫
  Appearance (Weight, Height) ≫
  Gender

- Cf. gender/sex IRL is crucial for evolutionary survival but in comparison, is not as important in the Pokéverse.

“Real-world” attributes

- Results suggest that sound symbolism occurs when it will be most useful in distinguishing attributes, features, and classes for evolutionary fit.
- Real-world attributes in this way affect the categories that condition phonotactic differences.

The problem with Pokémon

- Pokémon names are “designed” for each character
  - Game designers have in mind particular Pokémon attributes to highlight when choosing suitable names.
  - (Though such choices may not be overtly driven by sound symbolism concerns, and are balanced with extralinguistic concerns such as trademarks).
- Pokémon exist in a fictional world.

Case study 2.

Sound symbolism in baseball player names

Why baseball players?

- A real-world dataset that parallels Pokémon:
  - human players similarly have physical attributes of weight, height, and power statistics

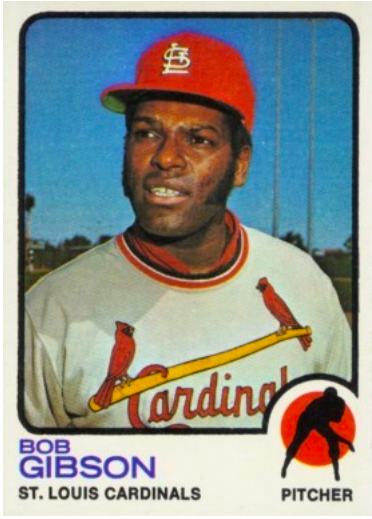
Major League Baseball (MLB) dataset

- *N* = 2557 baseball players
- Baseball player attributes and statistics
  - taken from live-ball era (1920–2017), of players with >450 total plate appearances
  - source: Lahman's Baseball Database (Lahman 2018)

Baseball player names

- Names in current case study
  - **registered** names – as officially registered in MLB
  - **given** names – as given at birth
  - **chosen** names – registered names that differ from given names
  - **nicknames** – baseball-specific nicknames (from a Wikipedia list, *n*=324)

given name	***Robert***
registered name (and **chosen**)	***Bob***
nickname	***Hoot***



Baseball player names

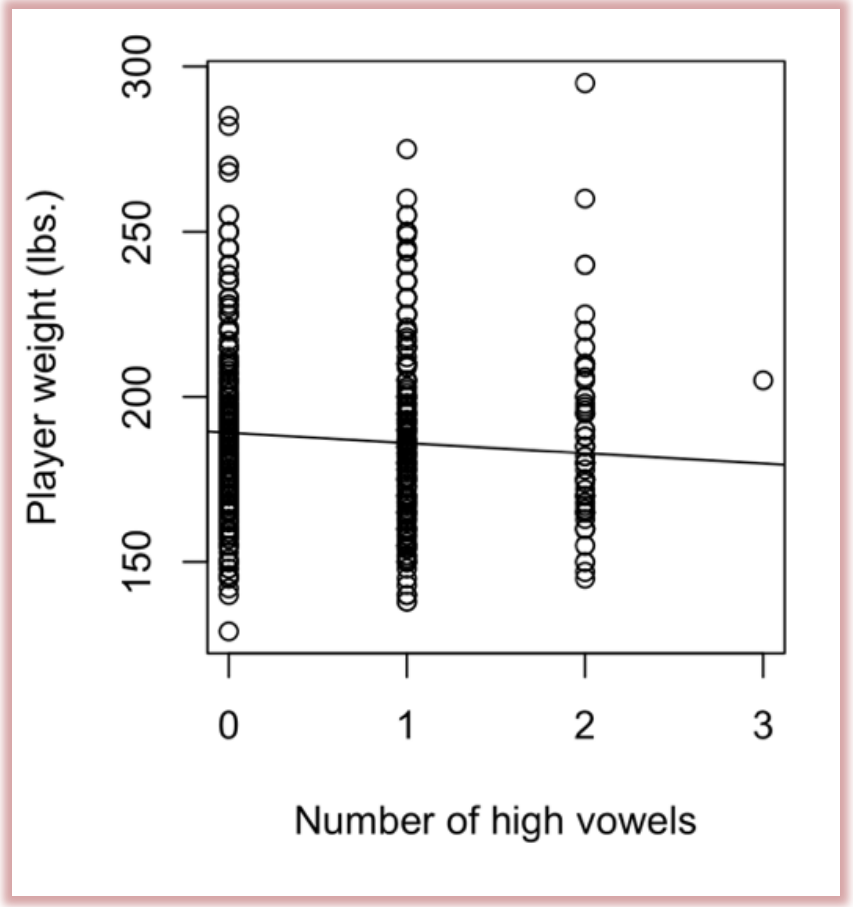
- Phonological features investigated: ~ Pokémon study
  - Prosodic factors
    - segments
  - Vowel quality
    - height
    - backness
  - Consonant quality
    - voicing
    - sonorant
    - POA: labial, alveolar, velar

Baseball player attributes

- Weight
- Height
- Power:
  - *slugging percentage* (~hitting power)
  - *batting average* (~plate discipline & accuracy)
  - *batting average on balls in play* (BABIP ~hitting skill independent of plate discipline and power)
  - *on-base percentage plus slugging* (OPS ~overall skill at plate)

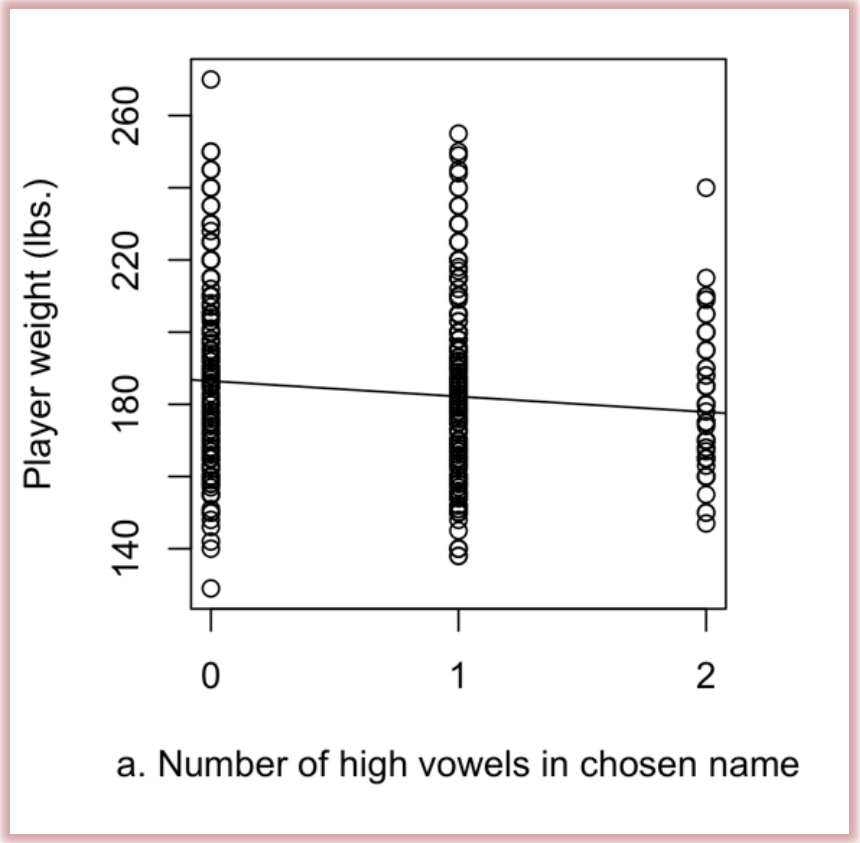
Results: high vowels

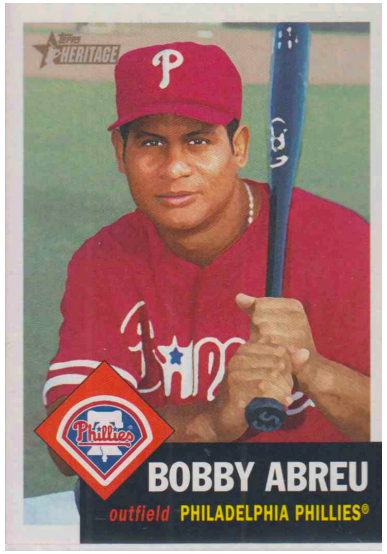
- In registered names, more high vowels negatively correlate with player weight (and height).



Results: high vowels

- In chosen names, more high vowels negatively correlate with player weight.
- No correlation in corresponding given names.

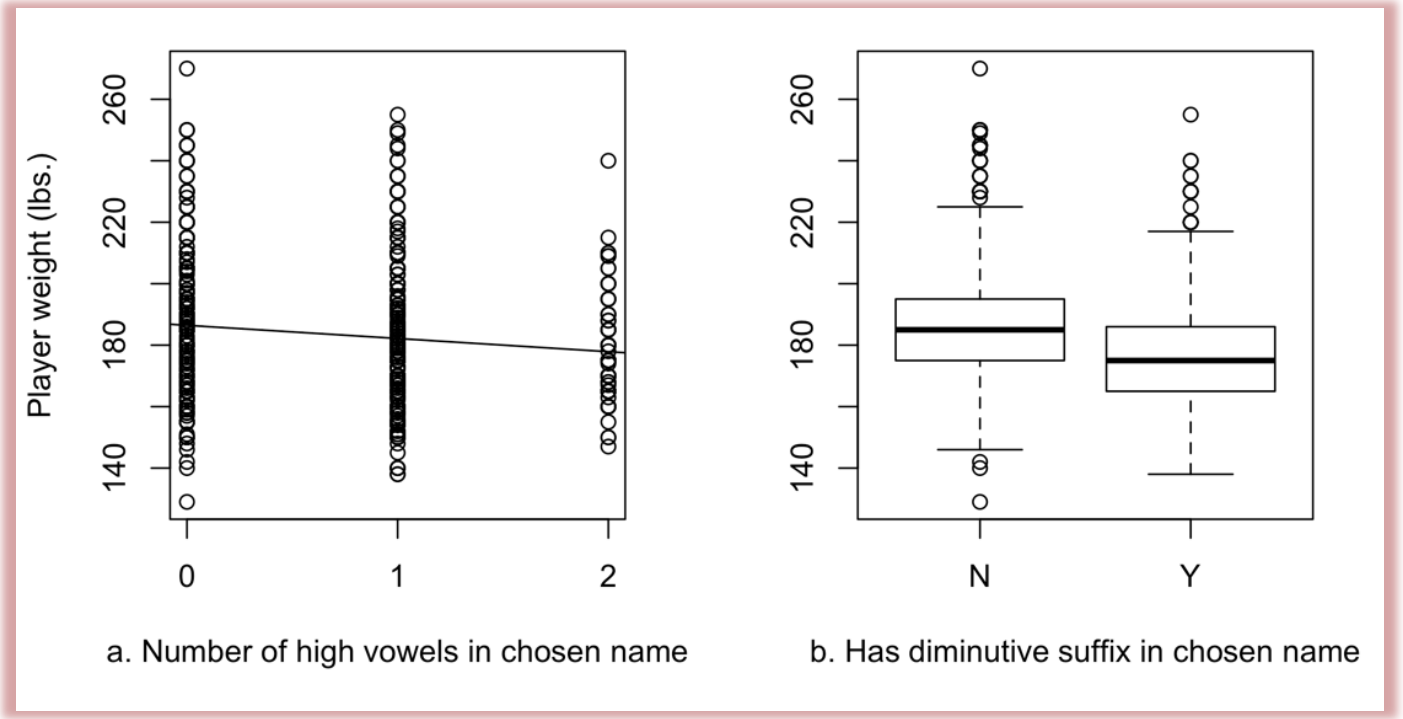




e.g., **Bob → Bobby**

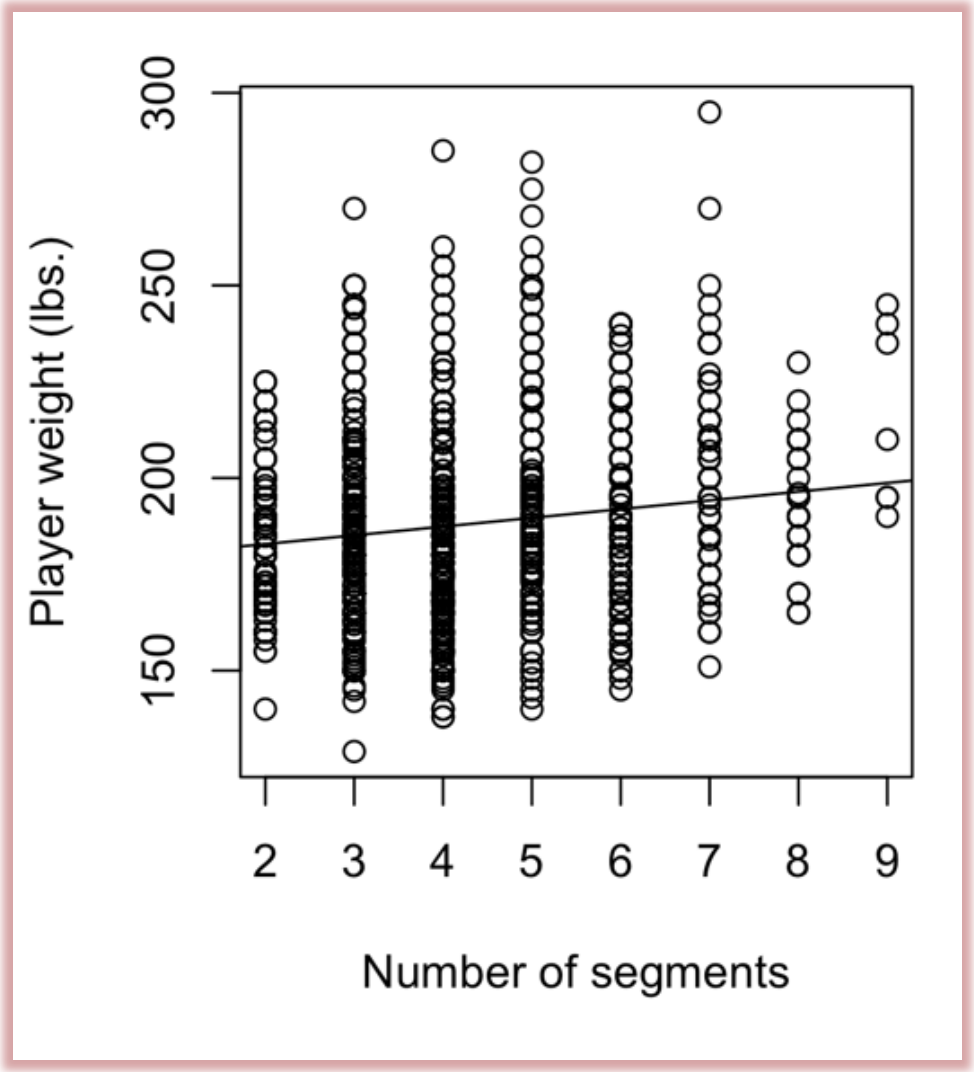
Results: high vowels

- Presence of a diminutive suffix (e.g., [–i]) actually is an even better predictor of weight than high vowels.
- High vowel sound symbolic effect is primarily driven by the diminutive suffix.



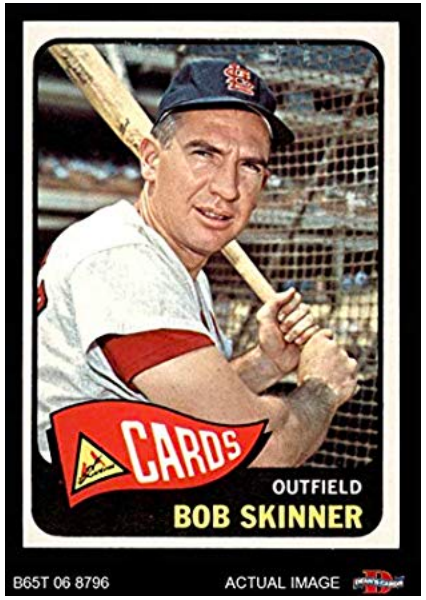
Results: length

- Longer registered names correlate with heavier players (and taller ones).

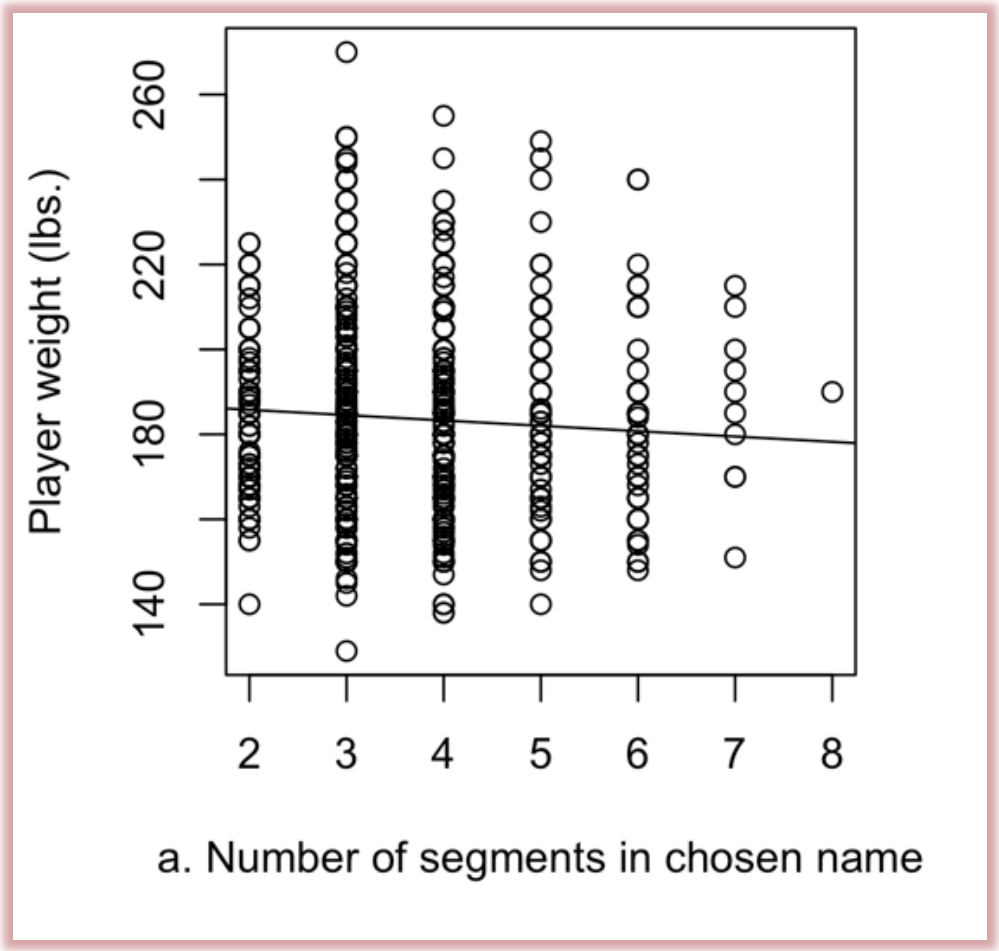


Results: length

- But, in chosen names, there's a negative correlation between name length and player weight.

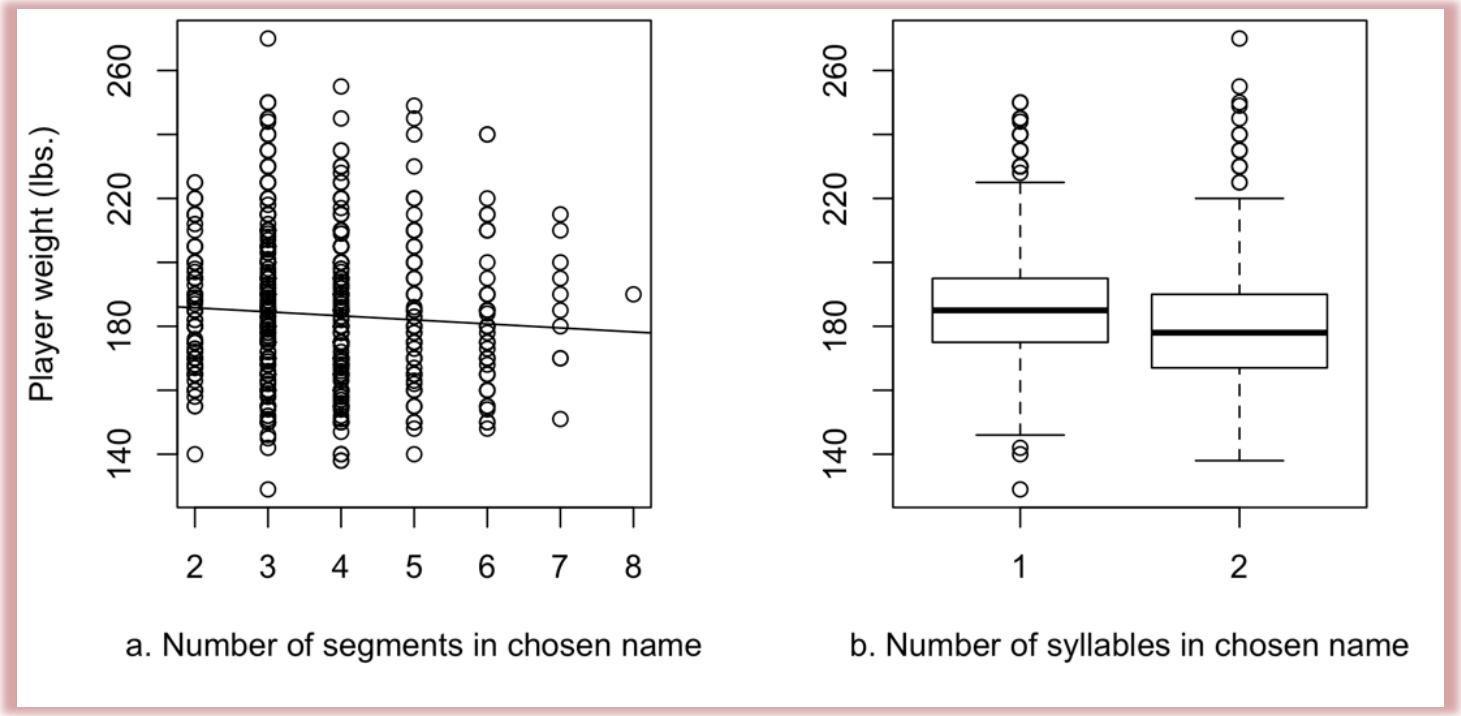


e.g., **Robert → Bob**



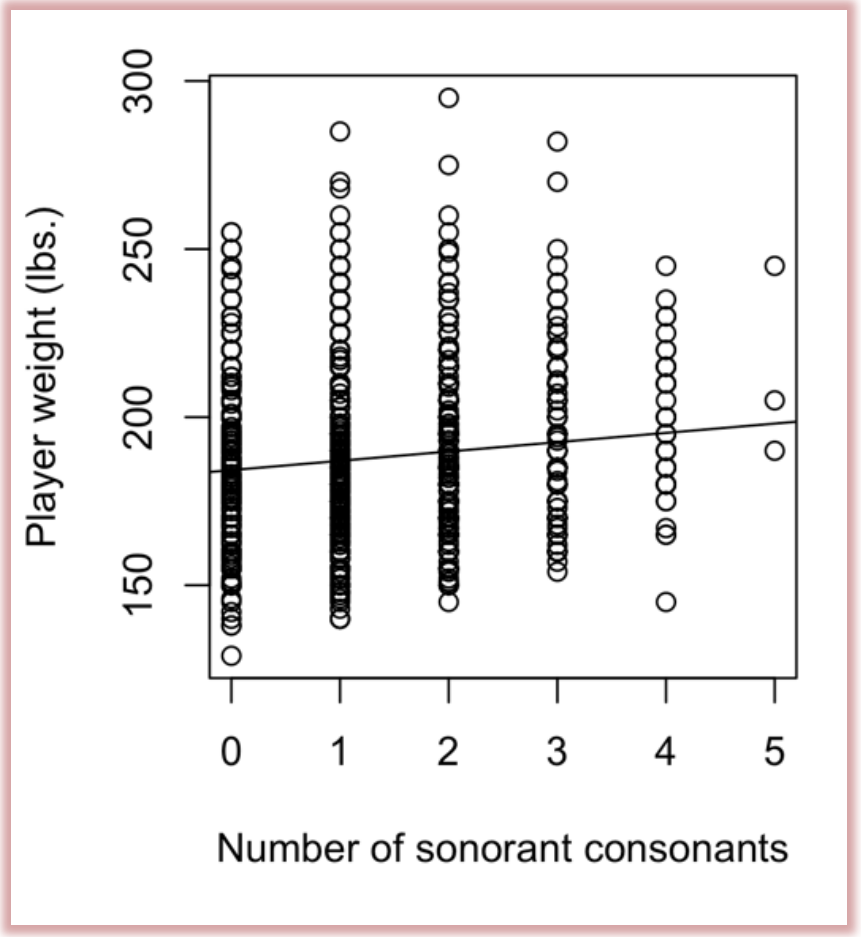
Results: length

- But, in chosen names, there's a negative correlation between name length and player weight.
- Effect largely in 1- and 2-syllable names.
- Truncation → heavier players.



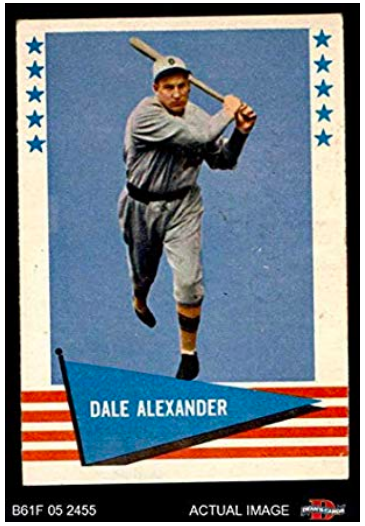
Results: sonorant consonants

- Registered names with more sonorant consonants correlate with heavier players.



Results: sonorant consonants

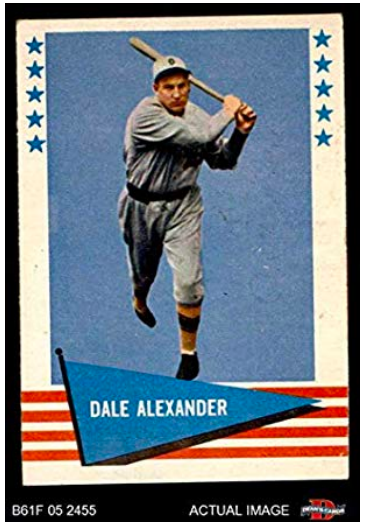
- Effect doesn't stem from a regular hypocoristic function in English (there are none that add sonorant consonants).
- Sonorant effect comes from chosen names where players choose to use their given middle name instead of their given first name.



e.g., **David Dale Alexander→ Dale Alexander**

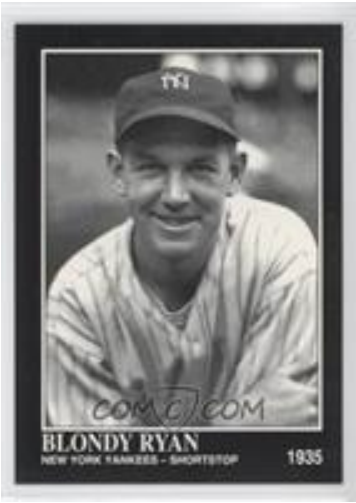
Results: sonorant consonants

- Effect doesn't stem from a regular hypocoristic function in English (there are none that add sonorant consonants).
- Sonorant effect comes from chosen names where players choose to use their given middle name instead of their given first name.



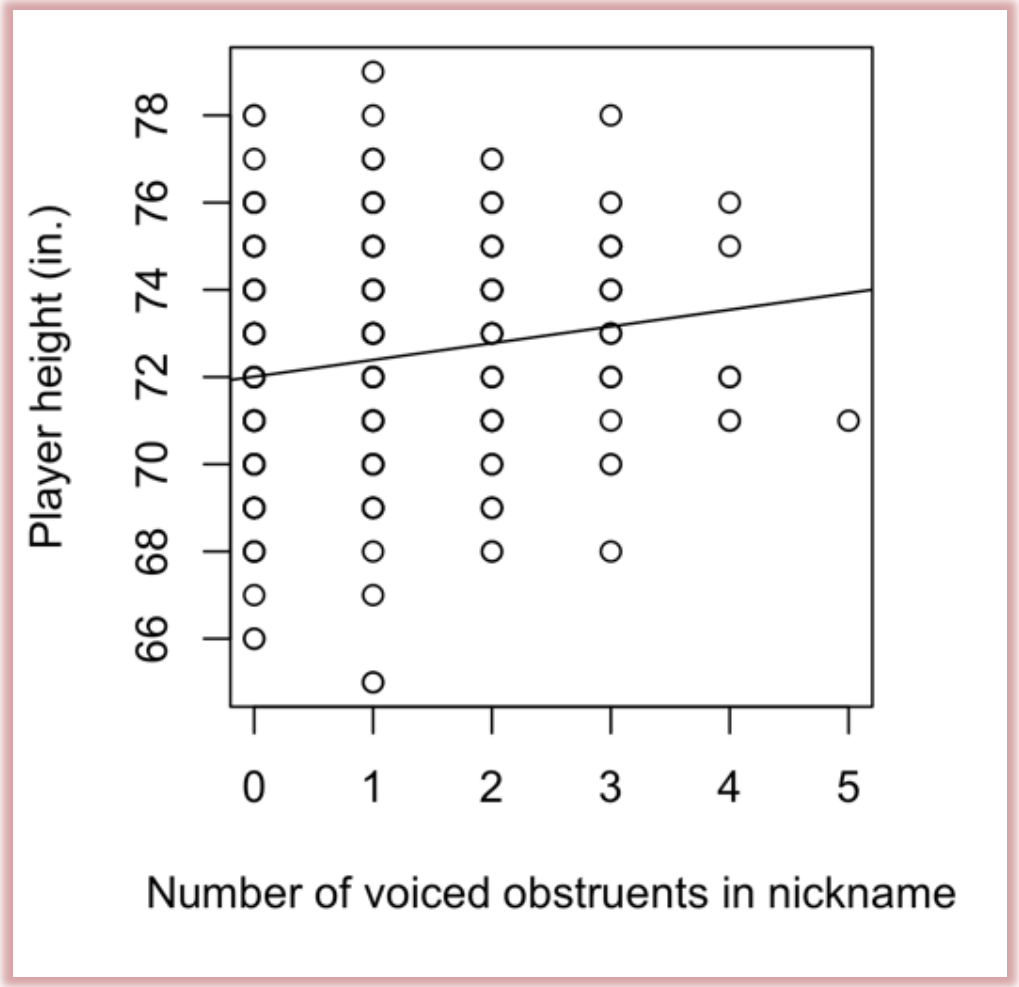
e.g., **David Dale Alexander→ Dale Alexander**

e.g., **John Collins Ryan → Blondy Ryan**



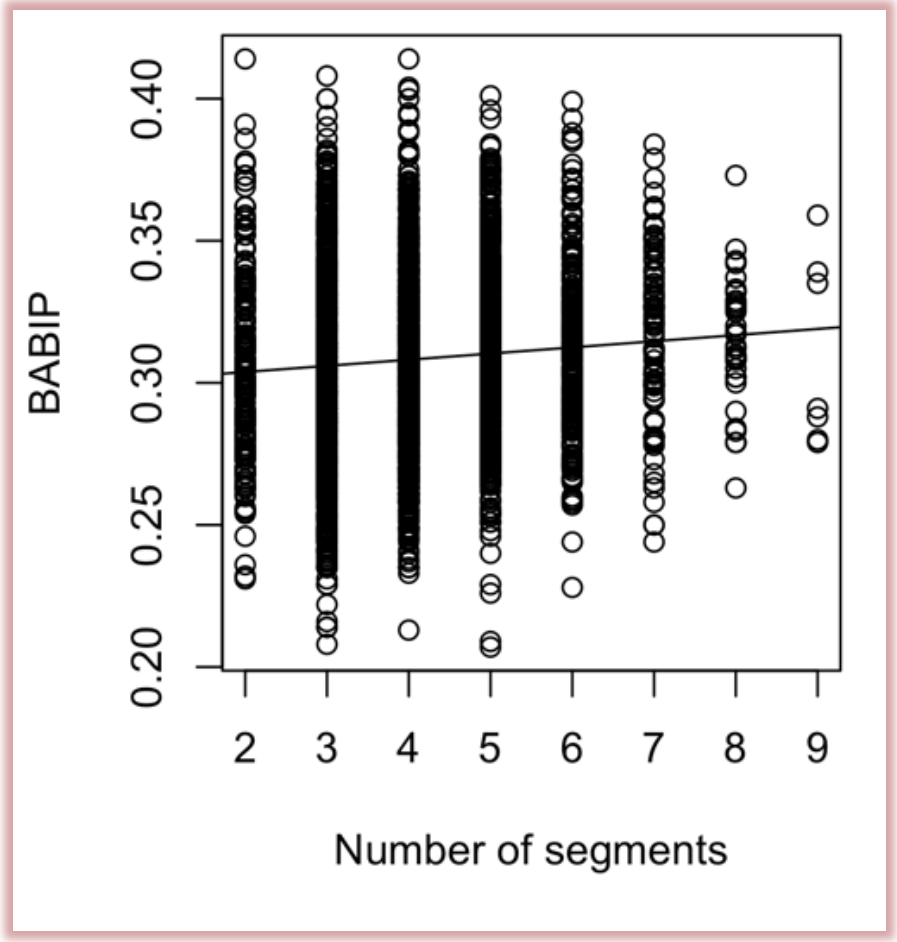
Results: voiced obstruents

- In nicknames, more voiced obstruents correlates with taller players.
- (Caveat: the data is rather sparse)



Results: baseball-specific attributes

- None of the baseball-specific attributes have phonological correlates except BABIP, with name length.
  - unclear why only BABIP (increases the chances that this is a spurious correlation somewhere)



Discussion: baseball

- While we find similar sound-meaning correspondences as with Pokémon, note that these are largely meted out by the existing English morphophonological system.
  - diminutive formation = high vowel sound symbolism
  - truncation = length sound symbolism
  - (choice of alternative name = segmental sound symbolism)
- suggests that there's an interaction/interface between sound symbolism and our "core" grammatical system.

Discussion: baseball

- For baseball players, physical attributes (e.g., weight) are most correlated with sound symbolic patterns, versus power (as with Pokémon), even though the latter is the most game-relevant.
- Potential explanation: Baseball players are human beings, who exist for many non-baseball-related purposes.
  - vs. Pokémon, which inhabit a constructed universe exclusively for gameplay-related purposes.

Discussion: baseball

- Nevertheless, the real-world attributes of baseball players do interact with formal phonological behaviours—when the diminutive applies, when to truncate, when to avoid a name if there's a hypocoristic gap.

Theoretical ramifications

Categoricity and gradience in the grammatical model

Lexically-conditioned phonology

Existing approaches:

- lexically-indexed constraints (e.g., Ito & Mester 1995; Pater 2000. 2009; Alderete 2009)
- strata (e.g., Kiparsky 1982, et seq.)
- cophonologies (e.g., Anttila 2002; Inkelas & Zoll 2005)
- sublexical grammars (e.g., Becker & Gouskova 2016; Allen & Becker 2015)

*** This is not an exhaustive list.**

Lexically-conditioned phonology

- Nearly all of these approaches assume crisp, discrete boundaries in category membership.
  - content word or function word
  - noun or verb
  - Latinate or non-Latinate
  - male or female

  - e.g., lexically-indexed constraints for the “expressive” lexicon (e.g., Alderete & Kochetov 2017; see also Kawahara et al. 2019)

The problem of category membership

- Not all sound symbolic patterning corresponds to crisply delineated category membership.
  - Categorical membership:
    - evolutionary stage
    - gender
  - A bit fuzzier?
    - weight
    - height
    - power

The problem of category membership

- Not a new issue in linguistics!
- See e.g., scale structure in semantics (e.g., Kennedy & McNally 2005; Kennedy 2007)

(1) alive vs dead
(2) tall

The problem of category membership

- How do our grammatical models handle gradient category membership if the existing mechanisms that we have assume categorical (i.e., full or none) membership?

The problem of category membership

- **Option 1.** Posit a(n infinite) number of categorical cuts relevant for phonology.
  - Computationally rather inefficient.
  - Ignores the scalar/gradient nature of many category types.
  - To some extent, we do need to be able to posit new categories, but it's been shown (statistically, at least) that new categories are formed only when there is sufficient evidence to do so. (e.g., from psych: Ahn & Medin 1992; from stats: Burnham & Anderson 2002; for morphosyntactic categories: Shih 2018)

The problem of category membership

- **Option 2.** Allow gradience in the category structures that the phonological grammar operates over.

The problem of category membership

- **Option 2.** Allow gradience in the category structures that the phonological grammar operates over.

- Scaling and gradience have been shown to be necessary in many parts of phonological grammar: (Hsu & Jesney 2017 for most of this list)
  - Trigger-target distance in vowel harmony (Kimper 2011)
  - Morphosyntactic distance (McPherson & Hayes 2016)
  - Lexical frequency (Coetzee & Kawahara 2013; Adams 2014)
  - Prosodic boundaries (Hsu & Jesney 2016)
  - Gradient symbolic representations over phonological elements (Smolensky & Goldrick 2016; et seq.)

*** This is not an exhaustive list.**

Gradience in category membership

- **Proposal.** Gradient symbolic activations on category membership.

Gradience in category membership

- **Proposal.** Gradient symbolic activations on category membership.

$$w_1(\mathbb{C}) + w_2(\mathbb{C} \times \mathbb{k}_i) + w_3(\mathbb{C} \times \mathbb{k}_j) + \ldots w_N(\mathbb{C} \times \mathbb{k}_x)$$

Each constraint $\mathbb{C}$ has a weight w_N for every category $\mathbb{k}$,
where each $\mathbb{k}_x \in [0, 1]$.

Gradience in category membership

- **Proposal.** Gradient symbolic activations on category membership.

$$w_1(\mathbb{C}) + w_2(\mathbb{C} \times \mathbb{k}_i) + w_3(\mathbb{C} \times \mathbb{k}_j) + \ldots w_N(\mathbb{C} \times \mathbb{k}_x)$$

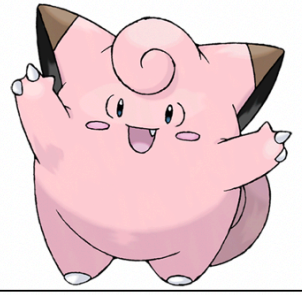
Each constraint $\mathbb{C}$ has a weight w_N for every category $\mathbb{k}$,
where each $\mathbb{k}_x \in [0, 1]$.

- $\sim$ multiple membership multilevel models
  (e.g., Hill & Goldstein 1998; Brown et al. 1998; Rasbach & Browne 2001)

Gradience in category membership

		$\mathbb{C}_{\mathbb{k}}$	
		1	$\mathcal{H}$
(1) input 1, $\mathbb{k}$=0.9	output A	–1*0.9	–0.9
(2) input 2, $\mathbb{k}$=0.2	output A	–1*0.2	–0.2

Gradience in category membership

		AVOIDREDweight	
		1	$\mathcal{H}$
(1) / weight=0.9 / 	pi.pi	–0.9	–0.9
(2) / weight=0.2 / 	pi.pi	–0.2	–0.2

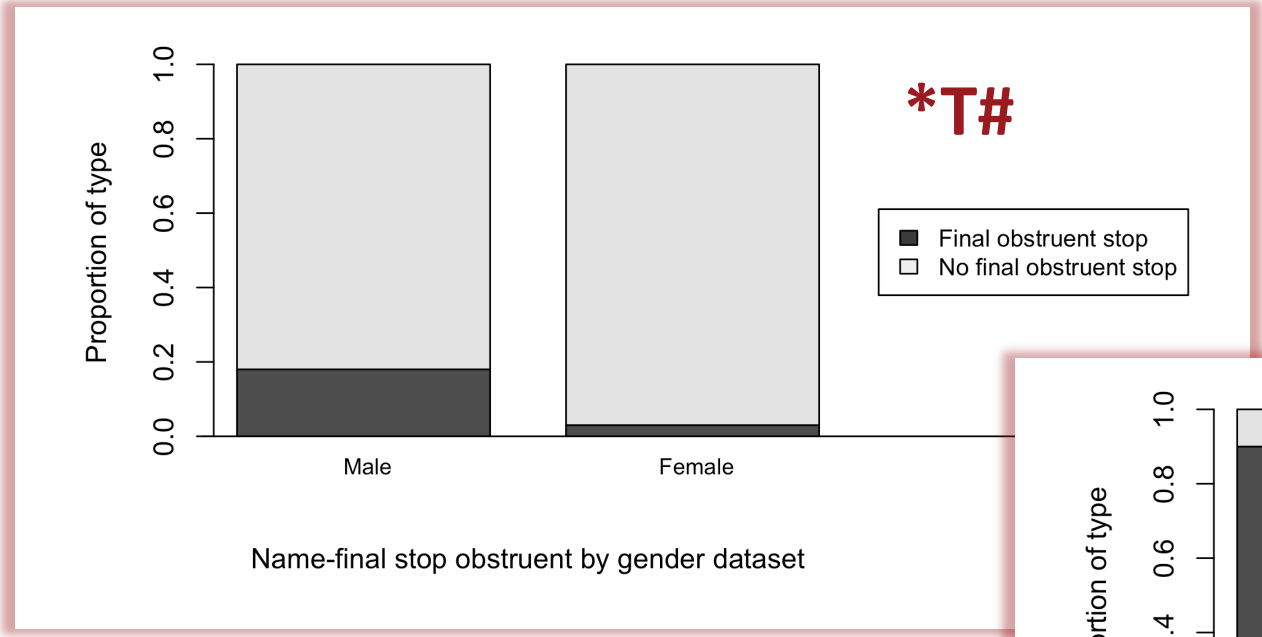
Gradience in category membership

		$\mathbb{C}_{\mathbb{k}}$	
		1	$\mathcal{H}$
(1) input 1, $\mathbb{k}$=0.9	output A	–1*0.9	–0.9
(2) input 2, $\mathbb{k}$=0.2	output A	–1*0.2	–0.2

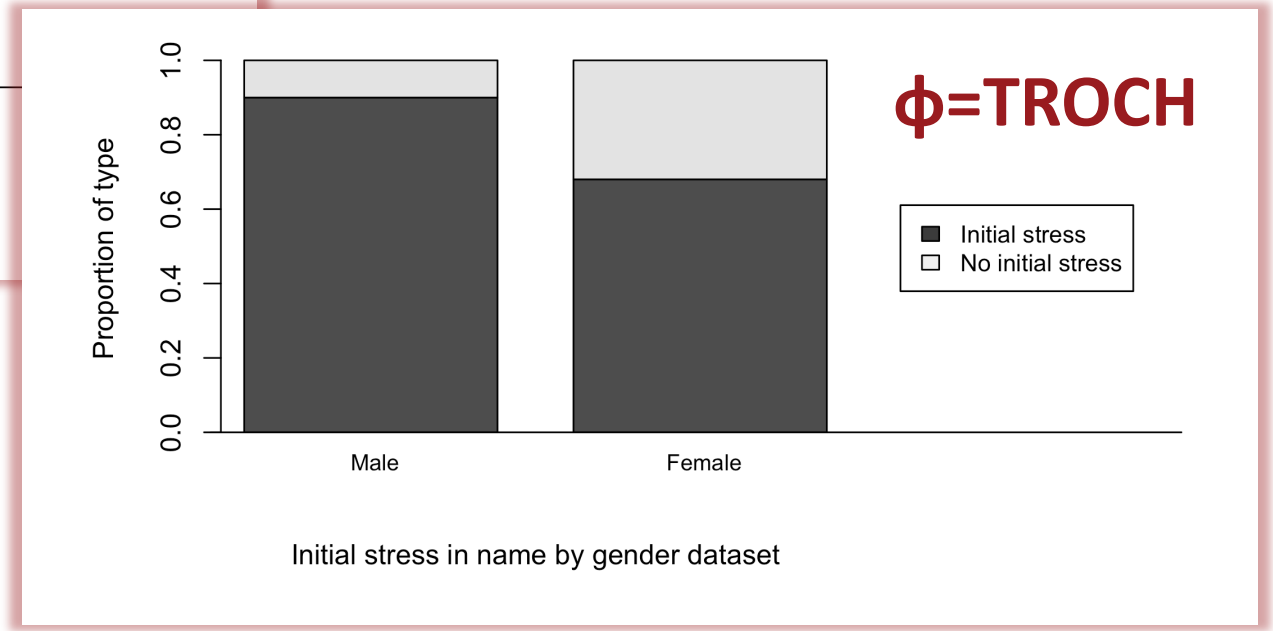
- Predicts that we should find phonological behaviours that fall between the categorical behaviours.

Toy Illustration

Male vs. Female names



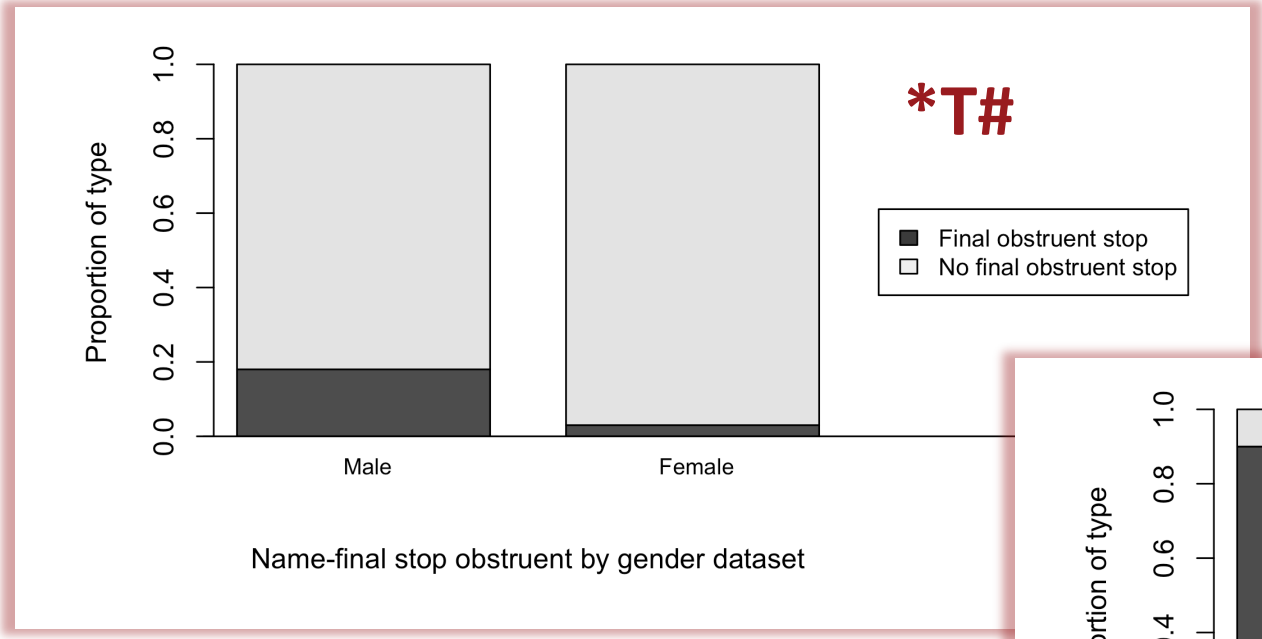
***T#**

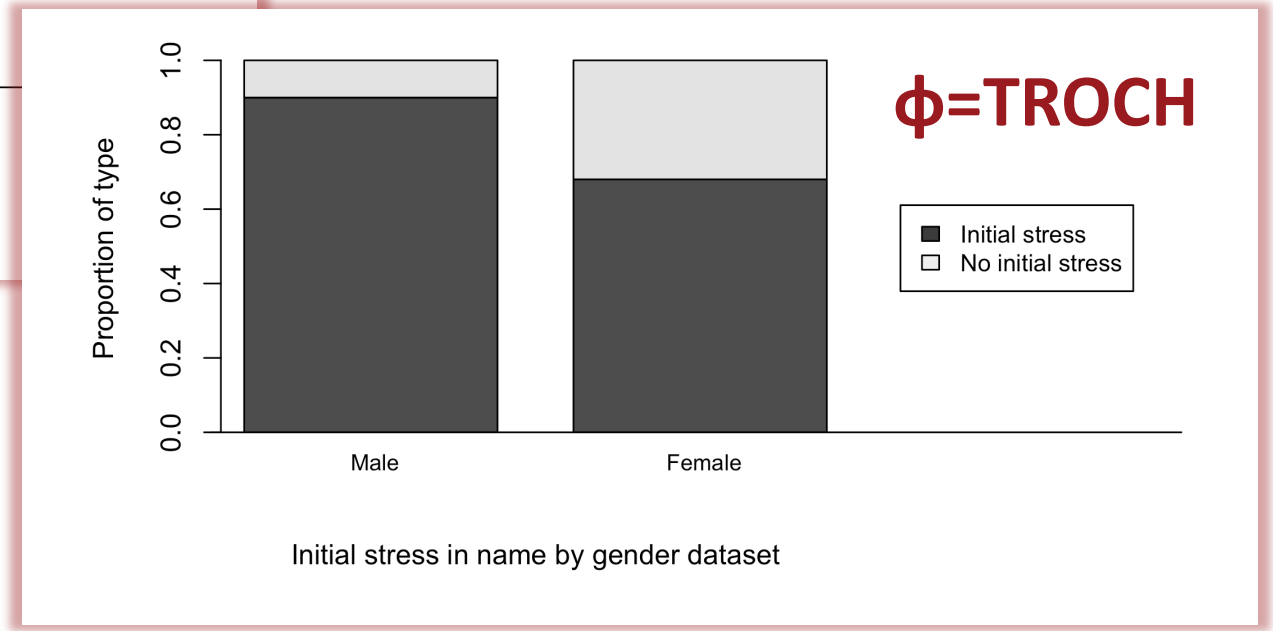


ϕ=TROCH

Toy Illustration

Male vs. Female names, and Unisex names





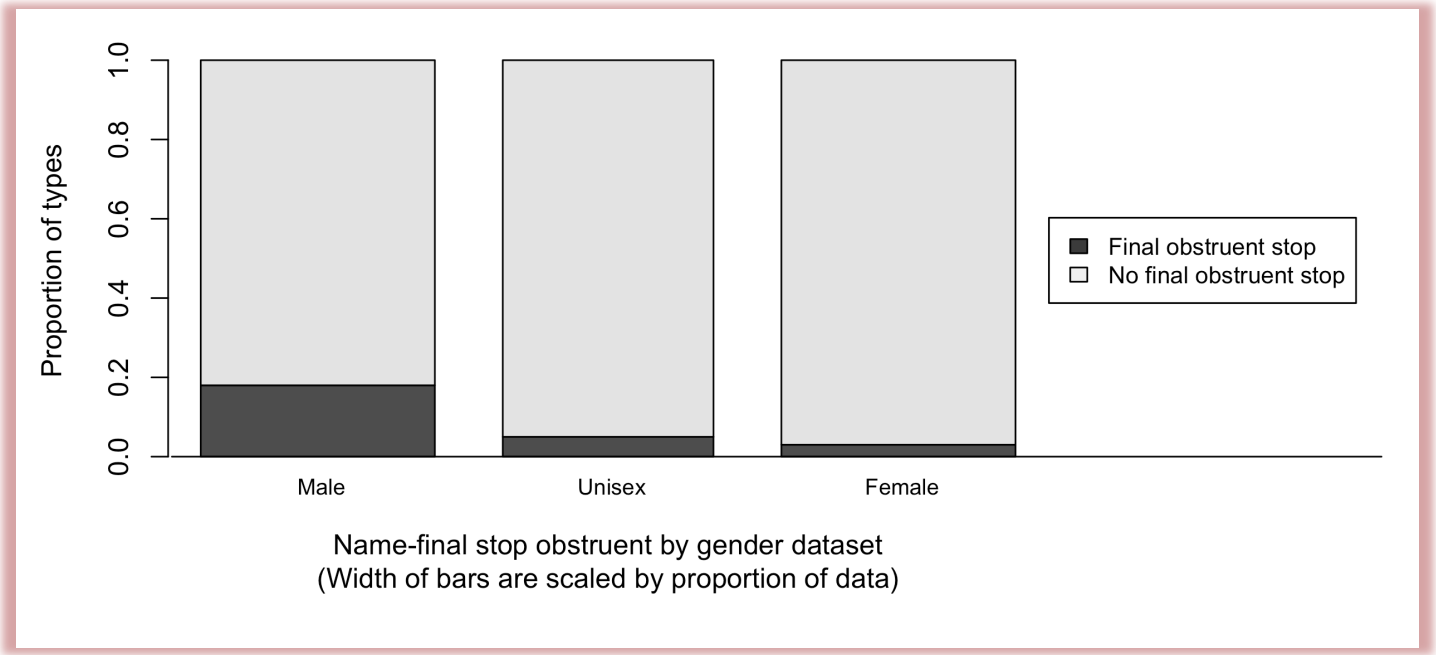
Toy Illustration

Male vs. Female names, and Unisex names

- Data:
  - Top 200 most frequent, most unisex names from the same dataset
  - ”Most unisex” = used for either male or female gender no more than 69% of the time. (i.e., as close to 50/50 as possible)

Toy Illustration

Male vs. Female names, and Unisex names



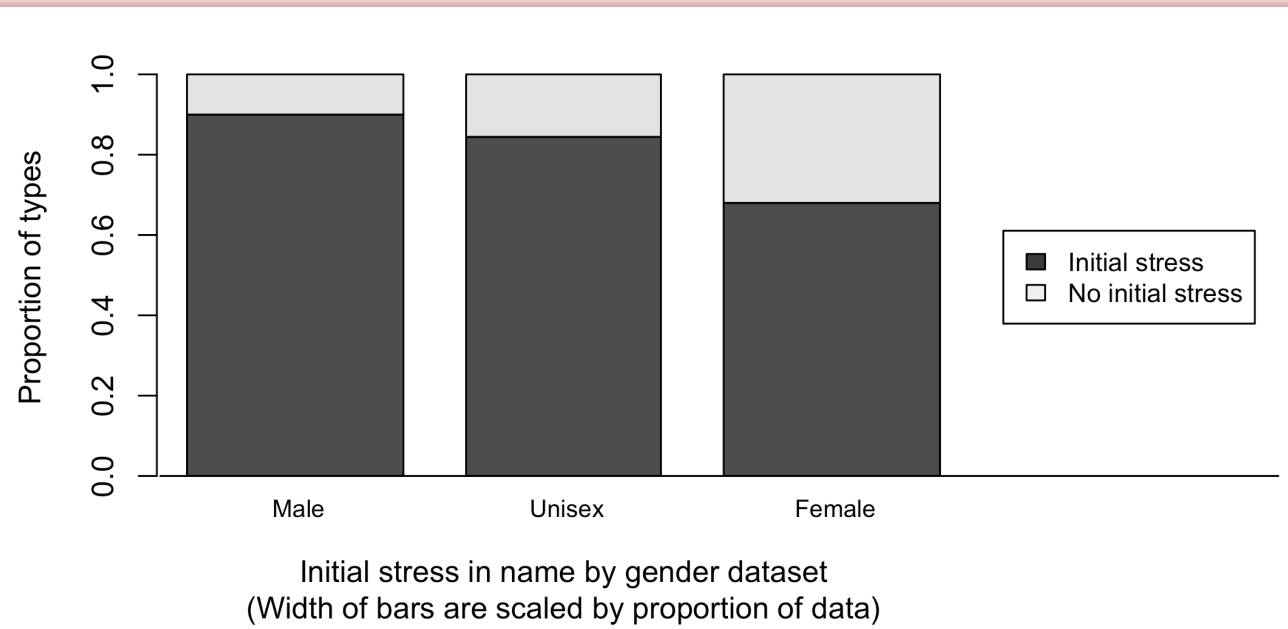
- ***T#**

Unisex names avoid final stop obstruents less than female names, but still ore than male names.

♀	⚥	♂
Elaine	Taylor	Albert
…[n]#	…[ɹ]#	…[t]#

Toy Illustration

Male vs. Female names, and Unisex names



- **ϕ=TROCH**

Unisex names are more likely to have initial stress than female names, but less likely than male names.

♀	⚥	♂
Elaine	Taylor	Albert
σ.ˈσ	ˈσ.σ	ˈσ.σ

Toy Illustration

Male vs. Female names, and Unisex names

♂ ♀ "Base" grammar

			TROCH♂	*T]#♀	WSP♀	TROCH	*T]#	WSP	FAITH-C	
			3	3	2	1	1	1	2	$\mathcal{H}$
(1) /CV.CVT/ ♂=1, (♀=0)	☞	a. ˈCV.CVT					–1	–1		–2
		b. CV.ˈCVN	–1			–1			–1	–6
		c. ˈCV.CVN						–1	–1	–3
(2) /CV.CVT/ (♂=0), ♀=1		a. ˈCV.CVT		–1	–1		–1	–1		–7
	☞	b. CV.ˈCVN				–1			–1	–3
		c. ˈCV.CVN			–1			–1	–1	–5
(3) /CV.CVT/ ♂=0.5, ♀=0.5		a. ˈCV.CVT		–0.5	–0.5		–1	–1		–4.5
		b. CV.ˈCVN	–0.5			–1			–1	–4.5
	☞	c. ˈCV.CVN			–0.5			–1	–1	–4

Gradience in category membership

- But do we need such gradient category membership outside of sound symbolic patterns?

Gradience in category membership

- Very likely: categories in lexically-conditioned phonology “proper” also exhibit similar behaviours.
  - e.g., auxiliary verbs (e.g., *can, could, might, must*)
    - like content words: host greater phonotactic contrasts
    - like function words: more likely to reduce (than full verbs)

  - To deal with cases like this, previous research has often posited that the relevant categories are more than just content versus function words:
    - e.g., 4 categories (Hirschberg 2004; Shih 2014, 2018; Anttila 2017); 10 categories (Altenberg 1987)

Conclusion

- Abstraction of form from meaning/concepts is no doubt an important component both in language and in linguistic theory.
- But: insights can be gained by examining the nature of the connections between form and meaning and their linguistic and extralinguistic contexts.
  - Especially so, for the design of our phonological systems.

Collaborators

Pokémon

Jordan Ackerman
University of California, Merced

Sharon Inkelas
University of California, Berkeley

Jessica Johnson
University of Southern California

Shigeto Kawahara
Keio University

Rebecca L. Starr
National University of Singapore

Noah Hermalin
University of California, Berkeley

Hayeun Jang
University of Southern California

Darya Kavitskaya
University of California, Berkeley

Miran Oh
University of Southern California

Alan Yu
University of Chicago

Baseball

Deniz Rudin
University of Southern California

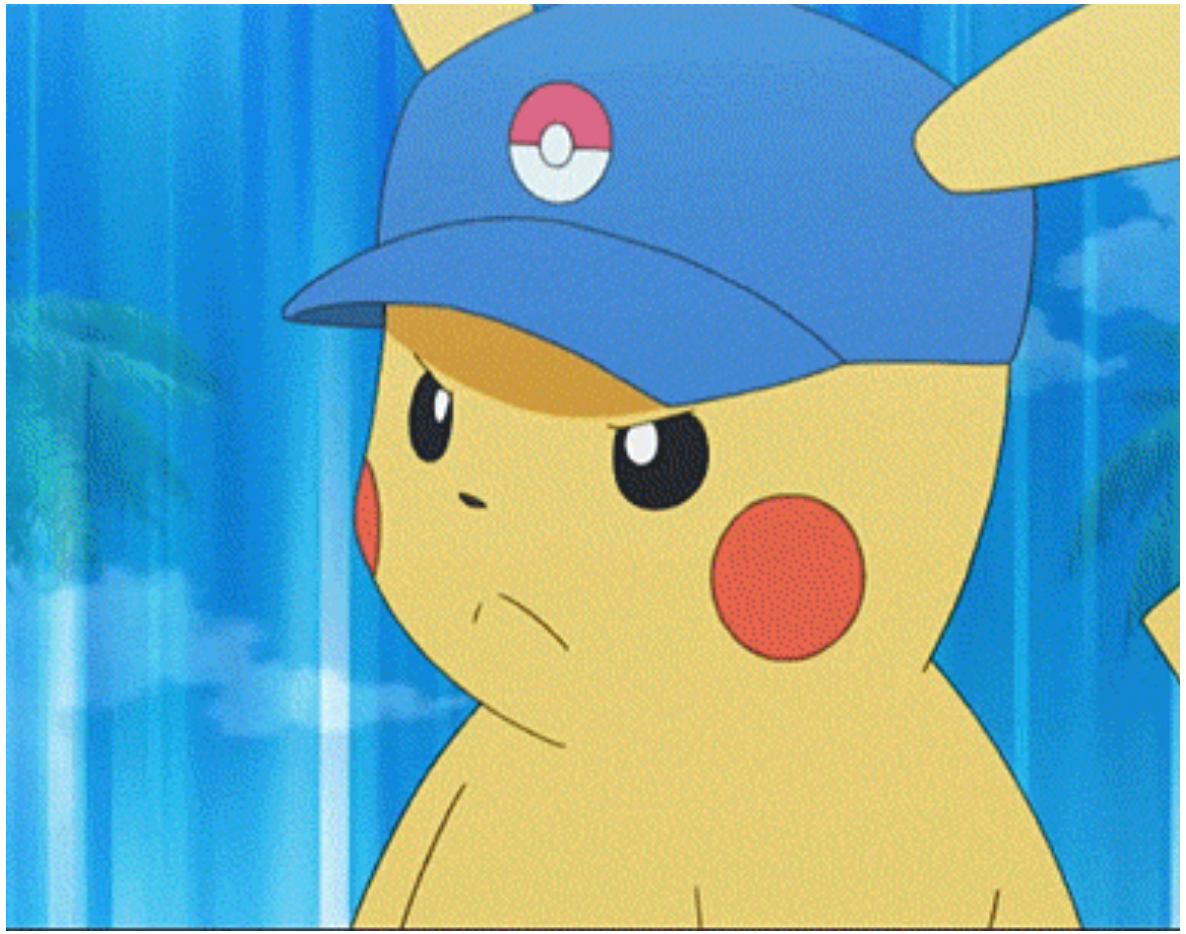
Pokémon RAs

Max Fennell-Chametzky, Seva Chapnin, Villi Gaaz, Katherine He, Alison Orgun, David Parker, Tianxiao Wang, Maria Whittle

Thank you

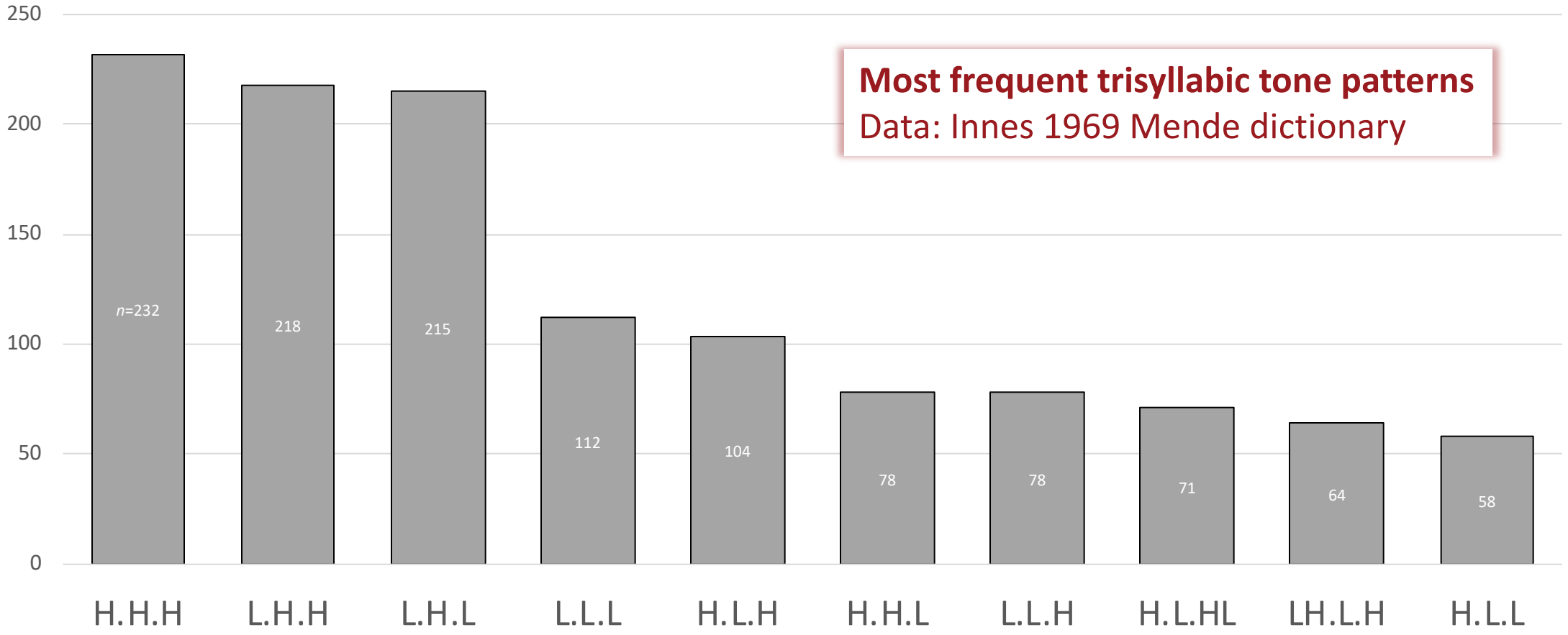
Acknowledgements to Darya Kavitskaya, Laura McPherson, Charlie O'Hara, Deniz Rudin, Brian Smith, Rachel Walker, and the USC PhonLunch and AMP 2019 audiences for discussion on various portions of this research.

Thank you to the organizers of NELS @ MIT!



A Classic Case

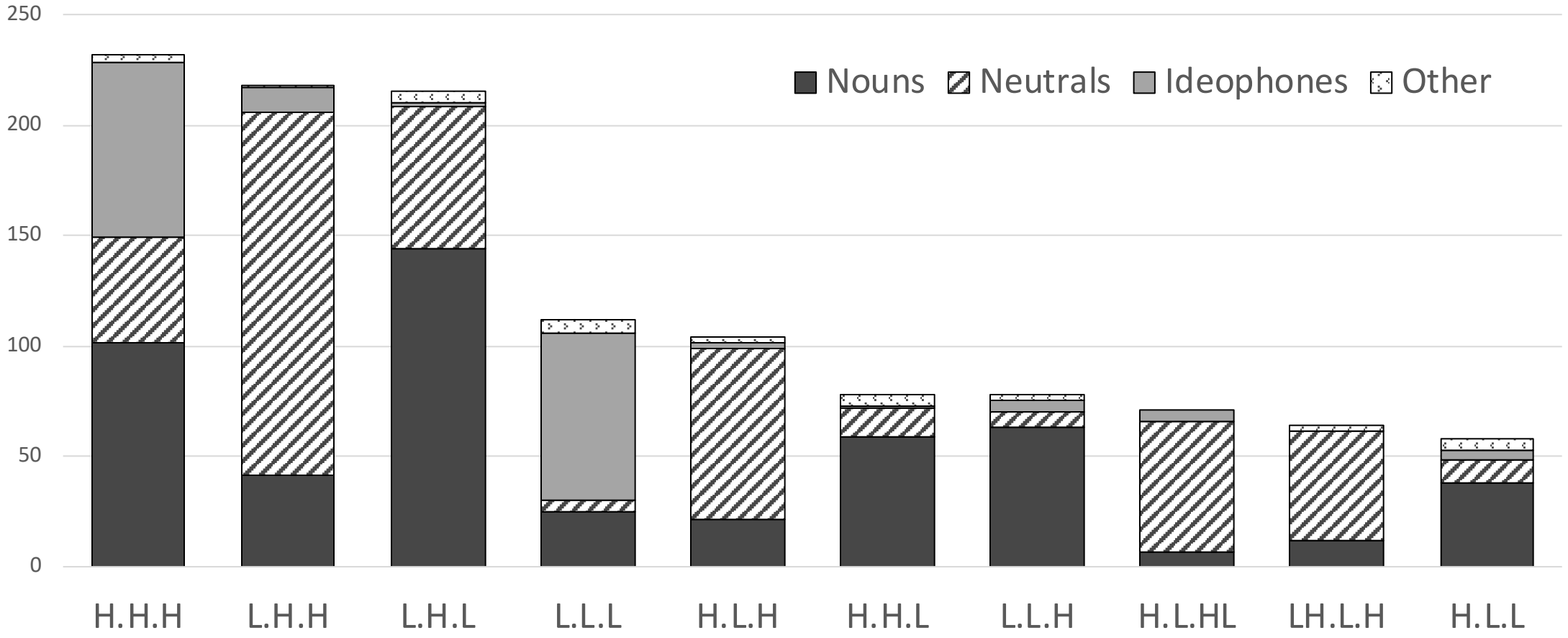
Tone Melodies in Mende



Most frequent trisyllabic tone patterns
Data: Innes 1969 Mende dictionary

A Classic Case

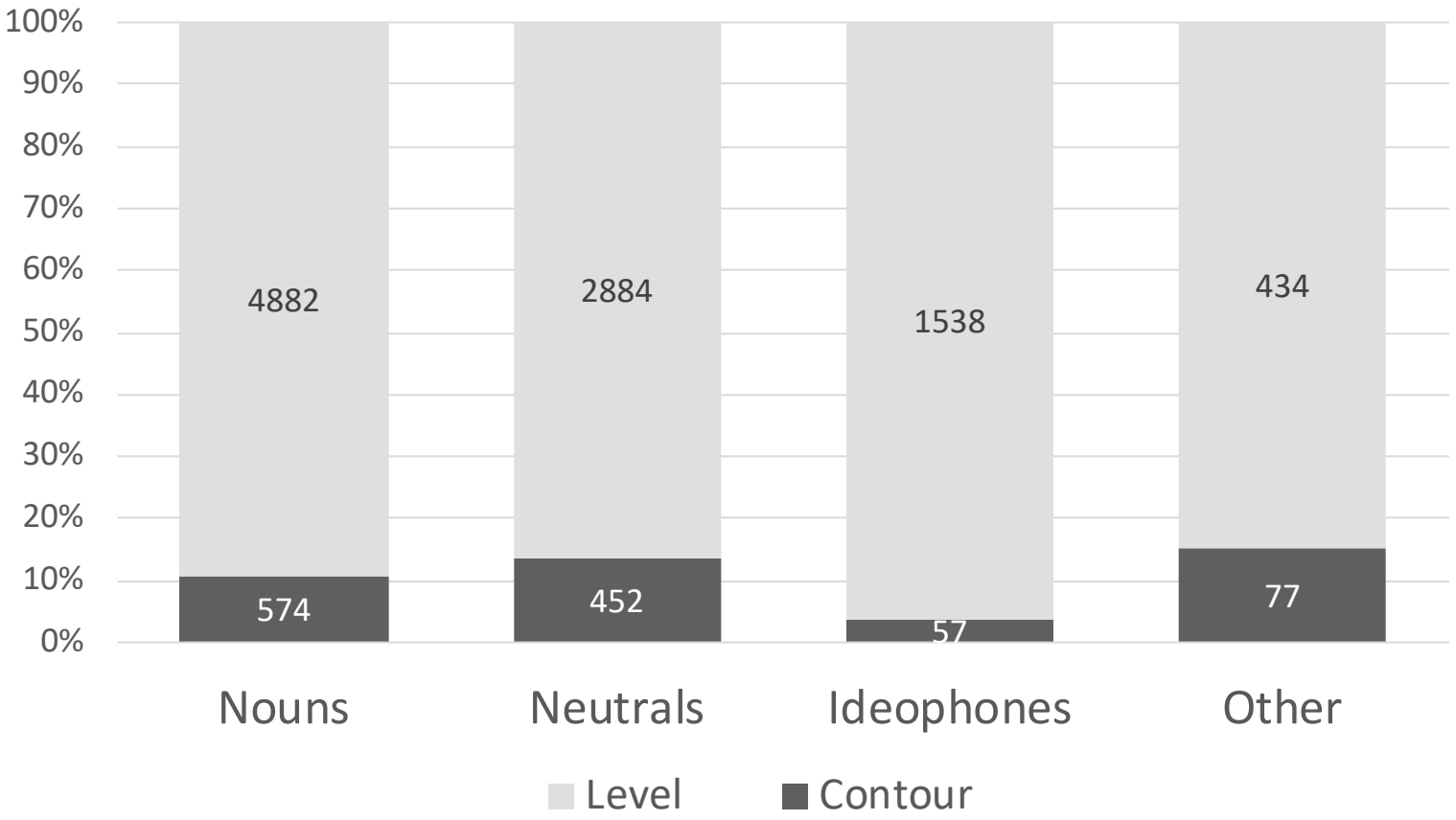
Tone Melodies in Mende



A Classic Case

Tone Melodies in Mende

Syllable-internal level (LL,HH) vs contour (LH, HL) tone patterns in Mende lexicon



- ***[αT][^αT] (AGREE)**

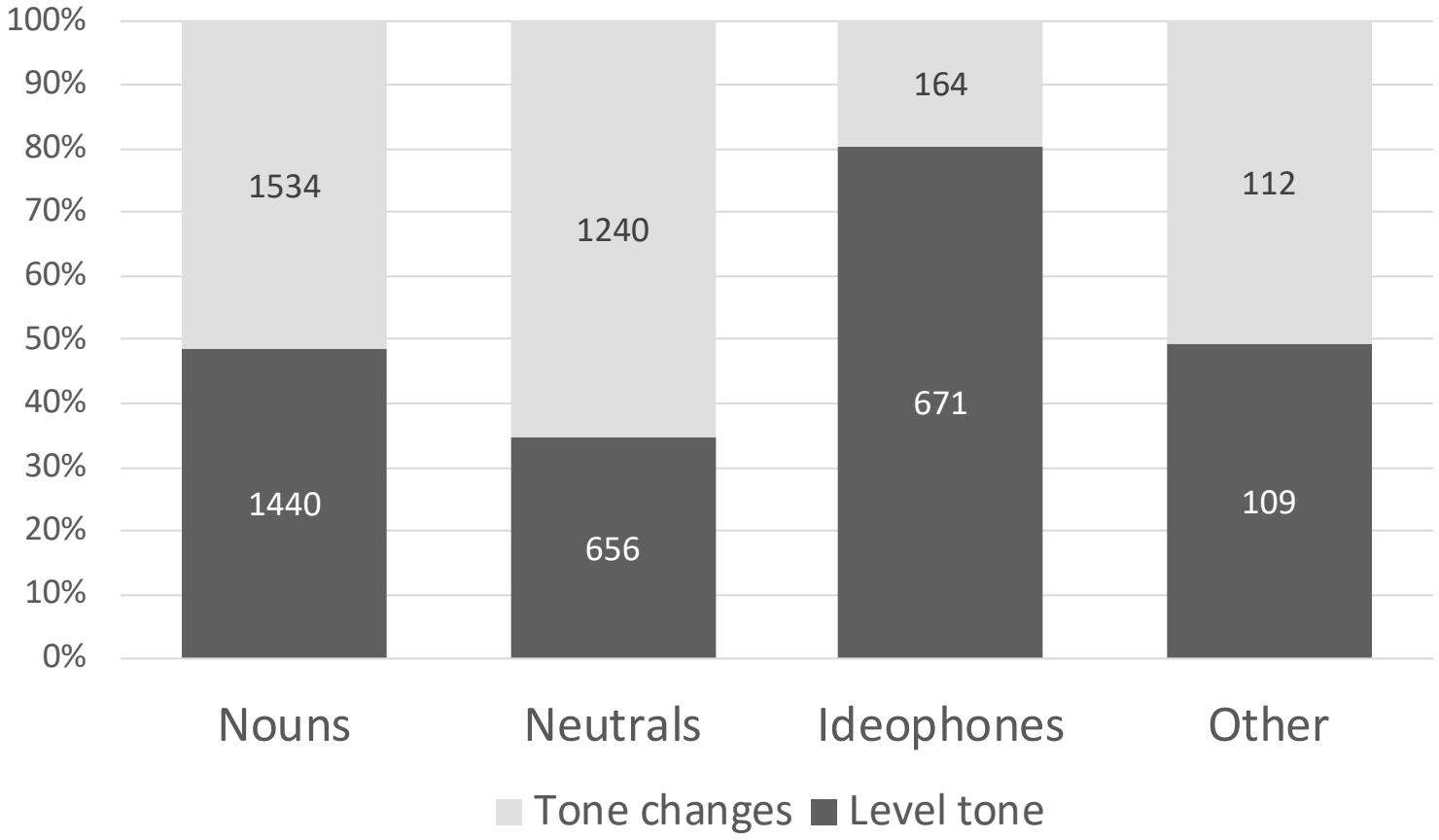
Avoid tone changes.

(Tones assessed at the subsegmental *q* level; i.e., 2 tones per each syllable, even for level tones: see Shih & Inkelas 2018)

A Classic Case

Tone Melodies in Mende

Syllable-internal level (LL,HH) vs contour (LH, HL) tone patterns in Mende lexicon



- ***[αT][^αT] (AGREE)**

Avoid tone changes.

- ***[αT]:$:[αT] (CHANGE@$)**

Avoid identical tones across syllable boundaries.

A Classic Case: Mende

Lexical conditioning in MaxEnt HG

Ideophone-indexed **Noun-indexed** **"Base" grammar**

			freq	AGREE$_{Ideo}$	CHANGE@$\$_{Ideo}$	AGREE$_{Noun}$	CHANGE@$\$_{Noun}$	AGREE	CHANGE@$\$	$\mathcal{H}$
				2	0	0	2	2	1	
(1) / σ.σ.σ / Ideo		a. LL.HH.LL	2	–2				–2		–8
	☞	b. LL.LL.LL	76		–2				–2	–2
		c. LL.LL.HL	1	–2	–1			–2	–1	–9
(2) / σ.σ.σ / Noun	☞	a. LL.HH.LL	144			–2		–2		–4
		b. LL.LL.LL	25				–2		–2	–6
		c. LL.LL.HL	30			–2	–1	–2	–1	–7

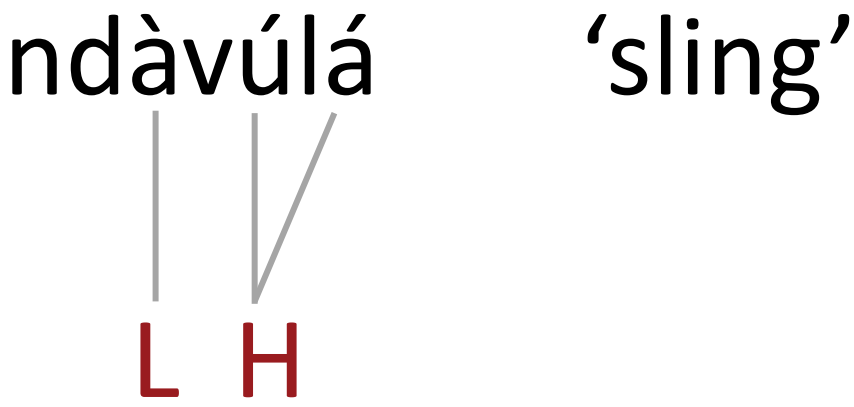
(Weights have been simplified for illustrative purposes here.
For a more complete analysis of Mende tonotactic patterns, see Shih & Inkelas 2015.)

A Classic Case

Tone Melodies in Mende

Traditional autosegmental view

- 5 underlying tone melodies
  - H, L, HL, LH, LHL
- Geometric association conventions
  - 1 ↔ 1, L → R



(Leben 1978; see also Hyman 1987 for similar melodies in Kukuya)

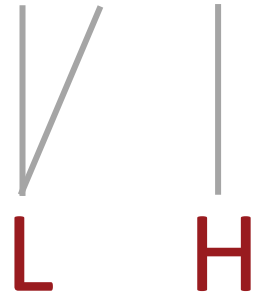
A Classic Case

Tone Melodies in Mende

Traditional autosegmental view

- Many surface patterns in the Mende lexicon deviate from the Autosegmental Phonology predictions. (Dwyer 1978; Conteh et al. 1983; Zoll 2003; Zhang 2007)

lèlèmá ‘praying mantis’



gbágbɛ̌mà ‘sensitive plant’